Event detection in social media: An analysis of Tweets from the Turkish Gezi Park Uprising

MASTER'S THESIS

Social media analytics, advanced statistics

Master's Program
"Management, Communication & IT"
Management Center Innsbruck

Thesis Advisor:

Dipl.-Ing. (FH) Kristian Hasenjäger

Author:

Karl Frederik Hunsmann

JGT17
1710576018

Submitted on:

August 12, 2019

Abstract

This paper takes an explorative approach to determine to which degree events that occur within in the context of a political uprising can be detected from social media posts with computational methods. This is based on two pillars: A comparison of systems and approaches that have been developed in previous research of this topic, and the application of a selected hierarchical clustering method to the case of the Turkish Gezi Park Uprising. The combined learning from these two steps is critically assessed, showing that event detection in the context of a political uprising is feasible, and that such a system can be used for situational awareness. Clustering methods are the most suitable choice for the detection of events by previously unknown categories. Combined applications of clustering and classification are highly suitable for the practical purpose of detecting events by known categories, whereas graph-based methods offer a more versatile approach for specific applications. On the use-cases side, intelligence services and humanitarian agencies are found to have an interest in event detection methods, which holds ethical implications for the deployment of such a system.

I. Table of Contents

II. List of figures

III. List of tables

IV. List of abbreviations used

AI	Artificial Intelligence
API	Application Programming Interface
CPU	Core Processing Unit

Df-idf$_t$	Document Frequency by Inverse Document Frequency and Time
GCP	Google Cloud Platform
GPU	Graphic Processing Unit
IDF	Inverse Document Frequency
JSON	JavaScript Object Notation
ML	Machine Learning
MP	Member of Parliament
NER	Named Entity Recognition
NGO	Non-Governmental Organization
NLP	Natural Language Processing
PM	Prime Minister
POS	Part of Speech
SA	Sentiment Analysis
SMA	Social Media Analytics
SVM	Support Vector Machines
TDT	Topic Detection and Tracking
TF	Term Frequency
TF-IDF	Term Frequency by Inverse Document Frequency
URL	Uniform Resource Locator

1 Introduction

The possibility of gaining insights from publicly available social media data is receiving attention from both research and practitioners. The data is interesting for a number of reasons: It can yield insights about social trends, the development of a language over time, and into the ways humans are connected with one another across the globe. Firms may monitor the opinion of potential customers towards their products and brands, while other organizations – including non-governmental organizations (NGOs) and governments – are interested in monitoring specific events or developments which affect their goals. Social media has been used to predict stocks (Bollen, Mao, & Zeng, 2011), product sales (Gayo-Avello et al., 2013), election turnout (Franch, 2013), disease spread (Sadilek, Kautz, & Silenzio, 2012), and even to detect earthquakes (Sakaki, Okazaki, & Matsuo, 2010).

In the field of humanitarian work, response time to disasters can be reduced through the use of early-warning systems that rely on big-data from multiple data sources (Jongman, Wagemaker, Romero, & de Perez, 2015). For example, data generated by ocean buoys may be used to detect a tsunami; location data from telecommunications providers may give insights into the resulting displacement, and keyword mining in social media data may yield insights into immediate needs of an affected population. In the crisis context, it has been shown that social media platforms function as vehicles for social convergence (Palen, Vieweg, Liu, & Hughes, 2009), as well as for collective coping and sense-making during disruptive events (Gaspar, Pedro, Panagiotopoulos, & Seibt, 2016) – all of which implies that such events leave a trace in the data that can be detected using computational methods.

In this thesis, I seek to provide an introduction to contemporary research in event-detection, which is a specific application of topic mining in social media analytics. The Gezi Park Uprising, which took place in Turkey in 2013, shall be used as a case to validate the applicability of such an event detection approach to a crisis-related data set, within the context of another country and language. This case is suitable, as a large amount of Tweets were produced in Turkey during the relevant time frame of the protests (Social Media and Political Participation Lab, 2013) and because Twitter

played an important role for information dissemination during this uprising (Ozturkcan, Kasap, Cevik, & Zaman, 2017).

1.1 Problem Statement

From my own experience, I recall that it was easier to find ad-hoc information about the unfolding events and the local security situation on Twitter than in the local media, which was censored at the time. However, isolating accurate information by comparing multiple reports and making sense of footage for situational awareness can be described as a time-consuming activity. Considering that computational event-detection methods in Twitter have already been designed and studied in previous research, perhaps the task can be made simpler through the help of technology.

From a technical perspective it can be assumed that a), different solutions to the event-detection problem consist of various but largely similar sub-tasks – such as defining the objectives, preparing the data, applying algorithms, and scoring the results – and b), that the chosen solution in each of these steps influences the outcome and the quality of the results.

Therefore, it is not clear to which degree a successful end-to-end approach can easily be reproduced and applied to a new case. Also, it is unclear whether differences in the data, such as the context of the topic or the language of the Tweets, present an obstacle.

1.2 Purpose and Objectives

The purpose of this thesis is to shed light on the current state of Twitter event-detection research, and to discover challenges associated with the reproduction of such a method to a new case. Therefore, qualitative part of this thesis is concerned with a comparison of different methods that have been developed in previous research. This comparison is to culminate in the selection of the most suitable method for event detection in the context of the Gezi Park Uprising case.

The quantitative part of this thesis then follows an exploratory approach, in which I apply the chosen method to Tweets from the Gezi Park Uprising in order to enable a comparison of the obtained results to the first study. The focus of the case chapter is therefore on the process steps required to achieve the objective, the challenges encountered, the distinct solutions that were implemented, and the quality of the obtained results.

The outcomes of this overarching approach should be an introductory overview of the research in this field, a documentation of the steps required to apply the chosen methods to the case, the results, and a retrospective evaluation of the method – information which could be useful for both researchers and practitioners. With this information, conclusions can be drawn about the feasibility of event-detection with computational methods for situational awareness.

1.3 Research Questions

Considering the research problem and the purpose outlined above, this thesis seeks to provide an answer to the following research question:

To which degree can events which occurred during the Gezi Park Uprising be detected with computational methods applied to Tweets?

This question will be narrowed down further according to the selection of the most suitable method for application to the case in chapter 2.5, through which the precise scope will be limited to the following:

To which degree can events which occurred during the Gezi Park Uprising be detected with [method name] applied to Tweets?

As an event definition, the definition provided by the Topic Detection and Tracking (TDT) project will be used, which describes an event as "something that happened at a specific time and place with consequences" (Allan, 2002). The success of the applied method will be evaluated first by the same scoring method(s) outlined in the original

paper in which the chosen solution was presented to allow a comparison of results, and second, by comparing the detected events to news reports about the events of the uprising. This is because I wish to find out whether a hypothetical user of the created event detection system would have been sufficiently informed regarding the security situation and the development of the uprising on a day to day basis.

The "to which degree" aspect of the research question will be answered in a detailed analysis of the nature of events which were detected by the system, by comparison to the nature of events which were *not detected* by the system, but reported in the news.

Since the event definition states the qualification criteria of *consequence*, a pre-selection of events from news reports will be used in which different levels of consequence are evident – i.e. whether the event was a *major* event with consequences for the overall course of the uprising, or whether it was a *non-major* event which had local consequences (eg. on individuals) but not necessarily on course of the uprising. This label assignment will be based on my own judgement and expert knowledge of the uprising, as I was living in the country at the time. In order to compensate for the lack of objectivity, and in order to be able to give a quantitative indication of event-retrieval performance despite the threefold selection bias resulting from choosing reports in advance (newspaper editor; Google search; own selection), the evaluation method described above will be complemented by an analysis of the rate of informative content returned by the system.

1.4 Outline

In chapter two, the topic of event detection will be introduced from a theoretical standpoint: The literature review is to begin with an overview of the fields of research engaged in social media analytics, and how event detection fits in this scheme. Following a theoretic topic introduction, different approaches to solving this problem which were found in literature will be described and compared. This comparison serves as the basis for the choice of a single, suitable method for application to the case, which will be identified in the subchapter "Selection of approach" at the end of the theoretical part.

Part two will begin with a high-level introduction to the Gezi Park case, including a timeline of events reported events. This introduction will be followed by the documentation of the steps undertaken to apply the chosen approach to the case. As a guiding framework for the case work, I have chosen to follow the CRISP data mining approach (Wirth & Hipp, 2000), which offers a six-phase process for data analytics projects. The CRISP method is suitable as it offers a more detailed framework to the general Capture – Understand – Present paradigm for social media analytics (Fan & Gordon, 2014):

- *Business Understanding*. The method previously selected for application to the case will be presented in more detail, which will be followed by a clear specification of the data mining goals and how success is to be measured. This will be complemented by an initial assessment of assumptions about potential constraints, resources required, and an estimation of risks and costs.
- *Data Understanding.* The dataset will be chosen and acquired based on the data mining goals defined in the first phase. An initial data exploration is to yield quantitative information about the features and their qualities, which will be described and visualized to provide an overview of the raw data. The output of this phase should be a table of the features, an indication of whether they are relevant to the objectives, and likely transformations that will need to be undertaken.
- *Data Preparation.* In the third phase, the data set will be cleaned and transformed to produce usable features: Irrelevant data points and features with lacking quality will be removed. Features will be transformed as needed, including the detection and removal of irrelevant outliers in the data and the addition of new features. This process will be documented clearly, including the tools, methods and system environment used. The goal of this phase is to produce a relevant data set which is clean enough to allow computational processing with reasonable results, and transparent enough to ensure results can be interpreted accurately.
- *Data Modeling.* In this phase, the chosen modeling approach will be applied step by step, including preliminary findings and a documentation of challenges that required an adaptation of the original approach – should this be necessary.

- *Evaluation.* In phase five, the results will be reviewed according to the previously defined business objectives, in order to answer the research question.
- *Deployment.* This final phase of the CRISP model will be replaced by a *review* of the approach, as deployment of this system for a practical purpose is beyond the defined research scope. The review phase will include a discussion of the implementation, the chosen parameter settings and their effect, as well as an analysis of potential for future improvement of the system.

The thesis will conclude with a summary of findings, an indication of the limitations of this thesis, and implications for research and practice regarding this topic. Finally, the last section includes the references, the affadavit and the appendix.

2 Literature review

Data mining is an application-driven research domain of computer science. While the term refers to the act of gaining knowledge from data (Han, Pei, & Kamber, 2011), executing a data mining task with computational methods implies both conceptual work – such as descriptive statistics – and the use of technical system layers. Therefore, data mining research is also related to other topics in computer science, including data storage, extraction, manipulation, and processing methods. The visual from Han et al. (2011) in figure 1 provides an overview of different data mining perspectives.

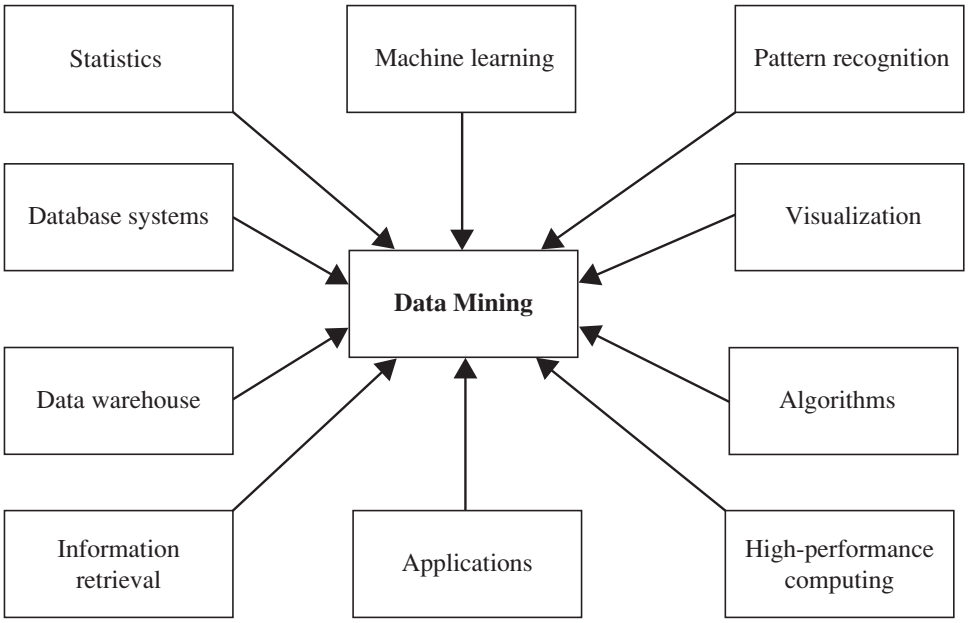


Figure 1: Research perspectives on the data mining topic;
illustration from Han et al. (2011)

Although data mining has been around for a long time, *big data* analytics only gained significance more recently through technical advances, i.e. the availability of greater processing power at a lower cost (Moore, 1965). As a result, the artificial intelligence (AI) domain of computer science has also undergone a revival, leading to the discovery of new technical approaches in machine learning (ML), search, and new applications for automation technology (Acemoglu & Restrepo, 2018). This has many implications: for example, certain industrial processes that previously required a human agent can now be

executed by a technical system. While automation is economically useful (Alitor & Salomons, 2018), it also holds potential for negative impact on society. For example, the storage of personal data and the use of AI in decisional systems pose considerable risks to a free and democratic society, resulting in the need for regulation (Scherer, 2016). Legislation and policy-making has to catch up to the expansion of technical capabilities. Therefore, political science, ethics, and philosophy are also engaged in AI and related research.

Data can be considered the building blocks of information, which in turn is defined as the aggregation of data to make "coherent observations" (Zins, 2007). Since a significant portion of all data on the web is stored in the form of unstructured text, and not in tables that describe clear relationships between elements, natural human languages present the dominant way information is encoded.

Natural Language Processing (NLP) is the research field at the intersection of linguistics, computer science and AI which seeks to bridge the gap between natural languages and computer systems. It can be described as a set of methods which make text machine-readable. These include low-level processing methods like detecting the boundary of a sentence, tokenizing words, part-of-speech (POS) tagging, morphological decomposition, and high-level methods such as the correction of spelling errors, and named entity recognition (NER) (Nadkarni, Ohno-Machado, & Chapman, 2011).

Boundary detection refers to the identification of sentence delimiters: For example, recognizing the difference between "Dr." and "word.". The first is a title, whereas the second marks the ending of a sentence. Tokenization refers to the conversion of substrings into a list of objects, for example "Alex walks down the road" into ["Alex", "walks", "down", "the", "road"]. POS tagging is assigning a predefined set of categories to the words: For example, tagging all nouns and verbs in a text. This low level task already presents a use case for neural networks. The same is true for morphological decomposition, which aims to reduce words to their stem or lemma to standardize word appearances: the stem is the mere reduction of a word to the shortest element which can be interpreted – which may or may not be a proper word by itself. For example, "going" may be stemmed to "go" and "destabilize" to "destabil". Lemmatization is similar, but the resulting smallest common denominator (the root word contained in the word) must

be a proper word: “destabilize” would remain unchanged, “friendship” would become “friend”, and “going” or “went” would both be mapped to “go” (own examples). In English, two of the most commonly used algorithms for morphological decomposition are the Porter Stemmer and the Lancaster Stemmer (Paice, 1994). NER refers to the recognition of named entities, such as people and organizations in text.

In chapter three, a computational data mining approach will be applied to Tweets, which places this thesis in the research context of Social Media Analytics (SMA), the discipline applying analytics to the data in the *blogosphere* of the web – which describes the user-generated portion of web content (Agarwal & Liu, 2009).

The blogosphere an interesting source of data to domains like social science and behavioral studies, which can be attributed to the sheer volume of user-generated content on the web and its representativity across geographic borders: 2.46 billion people or 33 % of the world population have social media accounts (STATISTA, 2018) (World Bank, 2018b), which translates to nearly 70 % of all the global population with internet access (World Bank, 2018a). On the other hand, SMA is relevant to domains like economics and marketing, which study the opinions about topics, how these develop over time, and which communities of consumers a product could be sold to (Melville, Sindhwani, & Lawrence, 2009).

Aggarwal (2011), who has surveyed the overall landscape of SMA research topics identifies 15 categories of work that has been accomplished in this field, covering both the structural aspects of social media as well as content-focused research. Data mining applied to social media *content* is one of these categories, whereas others include the analysis of the linkages in the network of users, how these evolve as the network grows, or the link prediction between nodes, the search and recommendations aspects (eg. the Page-Rank algorithm and applications of random walks), node classification, expert discovery, and community detection – to name just a few.

Since data mining is about uncovering hidden patterns in large volumes of data, machine learning (ML) approaches can be applied, where *supervised* learning refers to models that require training on labeled data, and *unsupervised* to those that can determine criteria for patterns and similarities autonomously (Pedregosa et al., 2011).

Although ML algorithms can be deployed with little effort, applying such a method in SMA requires a structured approach, involving data preparation steps such as preprocessing (eg. using NLP techniques to prepare the data), protecting the privacy of users, and estimating the impact of spam or noise in the given data set (Barbier & Liu, 2011). Thus, ML does not replace a thorough methodology.

Going deeper into different topics within data mining applied to the blogosphere, a multitude of topics have been explored, including blog classification (Bai, Wang, & Liao, 2009), topic propagation models and measurement (Gruhl & Guha, 2004), the identification of influential nodes (Agarwal & Liu, 2009), the detection of topics (Moon, Kim, Lee, & Oh, 2009), and sentiment analysis (SA) – which is the computational analysis of opinions, emotions, sentiments, and attitudes expressed in text towards an entity (Agarwal & Liu, 2009).

These categories can be considered the *building blocks* for more complex applications in event detection; therefore, some of the work in these fields is included here to provide a theoretic foundation to the next chapter.

One useful area of research from the domain of SA is sentiment classification: Using this technique, positive or negative sentiment can be detected, as well as vagueness in opinionated text, the languages used in a text, and the domain of a text body (Ravi & Ravi, 2015). However, the same goals can be achieved with different technical approaches – in this case, the machine learning approach, the lexicon-based approach, and the hybrid approach (Maynard & Funk, 2012). The machine learning approach makes use of supervised learning algorithms like the Naïve Bayes (NB) classifier and Support Vector Machines (SVM) (Medhat et al., 2014). Lexicon-based approaches, on the other hand, use dictionaries to correlate the usage of certain words to specific sentiments. Notably, such simple approaches can score better than machine learning techniques, given certain conditions (Nielsen, 2011). Finally, hybrid approaches can be applied, eg. a lexicon can be used to detect that an opinion is expressed, and machine-learning to determine the sentiment (Liu & Zhang, 2012).

Regarding the Twitter platform for applications in SMA, it is clear that platform-specific elements must be taken into account: The platform allows users to express

themselves in multiple ways that are not restricted to proper, English text. The 140 character-limit encourages ungrammatical text structures, and users frequently use non-standard elements like emoticons and colloquial expressions. For these reasons, a traditional lexicon-based approach may not be suitable (Saif, He, Fernandez, & Alani, 2016).

For the purpose of overcoming this informal text processing challenge, SentiStrength (Thelwall, Buckley, & Paltoglou, 2012) was developed, which relies on lexicons that support emoji, slang and "booster-words" (eg. absolutely, extremely) to output a weighed sentiment classification of Tweets. In a similar way, SentiCircles (Saif, Fernandez, He, & Alani, 2014) considers the semantic context of words by creating vectors for each term and its surrounding words. Then, the correlation between word occurrences is used to determine the weight of sentiment. This is to illustrate that multiple ways of achieving a similar goal may be possible.

Other studies in SMA have focused on what actually is being communicated on social media. Social media are a focal point of public discussion about current happenings, as anyone has the chance to express their opinion to a wider audience. The instant availability on any mobile smartphone device positions these platforms as vehicles of collective sense-making, and for coping with *disruptive* events.

For example, Gaspar et al. (2016) established a connection between social media sharing and social coping mechanisms, which they studied during a food poisoning crisis outbreak in Spain in 2011. Their finding was that a majority of shared content could be labeled according to different strategies for coping with crisis, such as information and support seeking, accommodation of the situation (self-management), escape, and venting of feelings of isolation and helplessness (Gaspar et al., 2016). Similar work on the 2007 Virginia Tech Shooting and the 2008 California Wildfires reveals that online social convergence is comparable to the social convergence that occurs in the geographic location following a disaster (Palen et al., 2009), when victims and relatives gather together to process what happened, and by-standers express support. So there is a correlation of social media activity to real-life events. From another perspective, this fact gained a lot of media attention during the events of the

Arab Spring, in which social media is reported to have influenced the birth of a mass movement (Khondker, 2011; Frangonikolopoulos & Chapsos, 2012).

Others have concentrated on hashtags and trending keywords to find that citizen journalism on Twitter plays a major role in information dissemination during a crisis (Öztürk & Ayvaz, 2018; Ozturkcan et al., 2017). Tweeters can take on different roles in this process (Varol, Ferrara, Ogan, Menczer, & Flammini, 2014). For example, an examination of retweet-rates was able to identify top influencers in the 2011 Egypt Uprising, and the way the geolocation of tweeters and retweeters affected propagation of information (Starbird & Palen, 2012).

Other studies have focused on trend detection in real-time on Twitter, for example the TwitterMonitor algorithm, which identifies bursts of keywords in a stream (Mathioudakis & Koudas, 2010). With such foundations in place, it is not a long shot to assume that information events by the Allan (2002) definition can be isolated using computational methods.

2.1 Event-Detection Methods

Detecting events in social media is a specific data mining problem closely related to trending topic and community detection; however, trending event information must be automatically distinguished from an abundance of non-event related content. This problem can be addressed in several ways, depending on how the data is modeled and represented conceptually before it is processed.

For example, the data associated with Tweets can be modeled as a document of features, such as the timestamp and the Tweet text. But it can also be represented as a graph in which the features are the nodes, and in which these nodes are connected with eachother by paths, which are called *edges*. Examples: “Tweet x”; “User a”; and “Tweet x was published by User a”. Machine learning methods like clustering and classification are typically applied to either document features (the *nodes*) in a graph, while search methods like random walks can take into consideration the edges in a graph. Such applications can also be applied in combination to yield specific insights, eg. clustering

followed by classification to determine the relevance of outputs. Thus, the same goal can be achieved through different data mining strategies.

Regarding overall event detection strategies, three categories of applications were found to be relevant in the surveyed literature: Simple *clustering methods*, combinations of *clustering and classification*, and *advanced approaches*. The advanced category describes a sample of approaches that are conceptually more complex than just NLP techniques combined with machine learning – including for example, graph partitioning and image recognition, which can only be presented on a very high level as detailing them would exceed the scope of this thesis. The following three subchapters provide an introduction to each category.

2.2 Clustering methods

The most frequently used technique in the surveyed event-detection research is *clustering*. Detecting events in Tweets can easily be framed as a clustering problem, as it can be assumed that if a multitude of posts refer to the same event, this would lead to a number of similarities in the data that can be detected and used as a criteria to group these posts.

Recalling the first section of chapter two, applying a ML model requires a clear methodology and preprocessing steps. In the case of clustering – an unsupervised learning approach – a *similarity metric* is required to allow pairwise comparison of the documents. Since this can be achieved most effectively with a numerical threshold, feature engineering is needed to transform the data into numbers. Clustering can be applied to different levels of information in a Tweet; however, for event detection purposes, the text, the timestamp, and the geolocation can be regarded as primary sources of useful information. Other features, such as the number of likes and retweets the Tweet received can not be clearly associated with event-relevance.

In order to model these numerically, the following transformations are proposed (Becker, 2011):

- *Text-based clustering*: Modeling word occurrence in a Tweet as a vector based on the total vocabulary in the dataset, and a weighting scheme.
- *Temporal clustering*: Conversion of the timestamp feature to a Unix timestamp – the number of seconds elapsed since January 1, 1970.
- *Spatial clustering*: Conversion of geolocation information or place references to coordinates.

Vectorization of text can be achieved by applying the Vector Space Model (VSM) (Hammouda & Kamel, 2004): Each document is represented as a *M*-dimensional vector with each dimension corresponding to a word in the total vocabulary of words that appear in the dataset. The terms are assigned weights according to their relative frequency in the document, which can be computed using simple term frequency (TF) (Manning, Raghavan, & Schuetze, 2009), inverse document frequency (IDF) (Robertson, 2004), or TF-IDF – which is the multiplication of both TF and IDF (Reed et al., 2006) (Luo, Chen, & Xiong, 2011). Pairwise document similarity can then be assessed by comparing the positions of the centroids (or average weight) of two vectors using a distance metric in the vector space. Several metrics have been proposed for this task, including the Euclidean distance (Heidarian & Dinneen, 2016), which is the straight line distance, and the cosine distance resulting from the angle between the two vectors (D'hondt, Vertommen, Verhaegen, Cattrysse, & Duflou, 2010).

In the case of Tweet timestamp comparison, similarity could be defined as $1 - \frac{|t1-t2|}{y}$, with y as the number of minutes in a year (Becker, 2011) assuming that all the Tweets were published within a one year time frame.

Analogically for geolocation comparisons, the similarity of two locations L1 = (lat1, long1) and L2 = (lat2, long2) can be modeled as 1−H(L1, L2), where H is the *Haversine* distance (Sinnott, 1984); a common metric for geographical distance.

Following these data preparation steps, a clustering algorithm can be applied to detect similarities at scale. Two very common algorithms to achieve this with social media data are the DBSCAN algorithm (Ester, Kriegel, & Xu, 1996) – a density-based clustering model that is not biased towards convex-shaped clusters, and K-Means

clustering (Steinhaus, 1956) – a method which seeks to partition the given observations into *k* clusters by mean distance to central points, resulting in a convex shape.

DBSCAN has been used to accurately detect large, local events in a noisy Twitter dataset by grouping location and hashtags co-occurrences (Ranneries et al., 2016). It has also been applied successfully to detect events in a real-time Twitter stream (Feng et al., 2015), but regardless of the data ingestion scheme it was shown to be generally outperformed by simple K-Means clustering (Yang & Rayz, 2018). However, the fact that the number of clusters must be specified in advance makes it unsuitable for detecting previously unknown events (Becker, 2011).

Becker therefore proposed using a threshold-based clustering technique which is able to handle a stream of new documents. Typically, hierarchical clustering algorithms are used for this task: In the agglomerative "bottom up" approach, each of the documents are initially assigned a cluster and then merged, whereas in the divisive "top down" approach the starting point is one single cluster which is then recursively split to build clusters (Berkhin, 2006).

However, Becker argued that hierarchical clustering does not scale well to a large data set, and that it cannot be applied to a stream of online data. Therefore, he proposes a "single-pass incremental clustering method" which considers each element in turn, and determines the suitable cluster assignment based similarity to clusters which were already formed – thus, beginning with a labeled data set and definition of a threshold for similarity: "The similarity score σ(d,c) is then defined as the similarity between document d and the centroid of cluster c. (…) This definition then avoids comparing document d against every document in cluster c." (Becker, 2011). σ was computed as the cosine distance between the vector centroid of the TF-IDF term weights of the Tweet and the average vector centroid of the cluster that it would be compared to. The resulting clusters of similar Tweets could then be compared and ranked according to a function f(Cluster, weight) to produce overall topic clusters. This concept is illustrated in figure 2, step c.

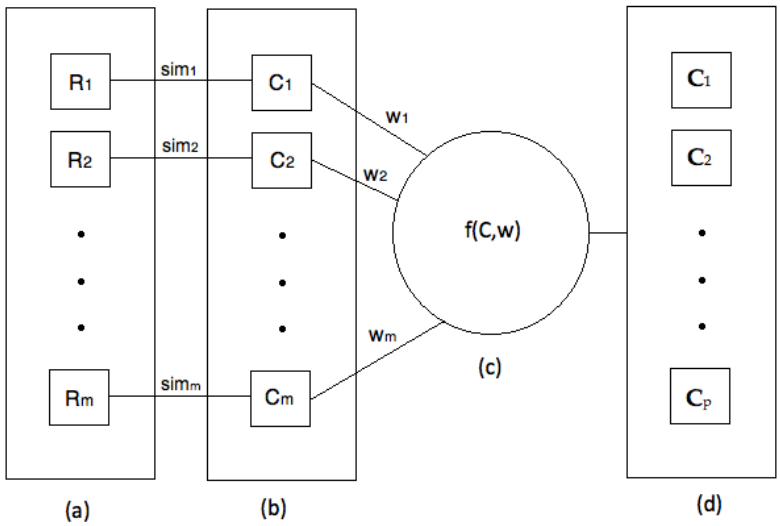


Figure 2: The clustering method proposed by Becker; illustration from Becker (2011)

However, just because two Tweets have similar content that does not necessarily mean the cluster is related to an event. *Ensemble methods* for clustering by multiple dimensions can be applied to achieve more targeted results, i.e. word, time and geolocation clustering applied together. Although the mathematical foundations have been laid for a multi-dimensional clustering approach with optimized similarity metrics (Domeniconi & Al-Razgan, 2009), the most suitable thresholds can also be determined empirically. This was demonstrated by Becker, who was able to accurately detect events once the thresholds were identified.

On the other hand, there are simpler ways to incorporate the time dimension in a clustering approach: Aiello et al. (2013) adapted the TF-IDF scheme to a weighting scheme that considers word occurrence by time frame, thus avoiding the need to cluster by several dimensions. Their $df\text{-}idf_t$ metric uses historical data to penalize such word groups that began to appear in the past and are still used frequently, but therefore do not define new topics.

Specifically, this was accomplished the following way: First, keywords were indexed and documents were organized by slots. Then, the document frequency was computed for each n-gram (recurring word or phrase) of the current time slot *i*. But as opposed to TF-IDF, the score is was penalized by the logarithm of the mean of its scores in the previous *t* time slots:

$$df\text{–}idf_t = \frac{df_i + 1}{\log\left(\frac{\sum_{j=i}^{t} df_{i-j}}{t} + 1\right) + 1}$$

Thus, a burst of similar keywords in multiple Tweets within a given time frame is ranked higher if the observation presents the first time that these keywords spiked. For greater relevance regarding *event* detection, they proposed boosting named entities – the proper nouns in the Tweet text – by applying an empirically determined boost factor of 1.5 if the n-gram contains a named entity (Aiello et al., 2013).

Ifrim et al. successfully applied this method in the following, integrated approach: Initially, the Tweets were divided by time windows of fiveteen minutes. Following "aggressive" preprocessing, a vocabulary list of uni-, bi- and trigrams which occurred in the time window was produced for each window, which was then modeled as a binary *tweet-term-matrix* – the basis for computing cosine distance as a document comparison metric. In this vocabulary, only words were included that exceeded a minimum threshold of *max(int(len(window corpus) ∗ 0.0025), 10)* appearances – meaning that for 10 000 Tweets the word should occur in at least twenty-five Tweets, whereas the absolute minimum requirement was occurrence in ten Tweets. The rationale of this filter was to ensure that clusters had to gather a sufficient amount of Tweets to be relevant (Ifrim, Shi, & Brigadir, 2014), thus influencing the *granularity* of detected events.

Then they applied hierarchical clustering, specifying a threshold of 0.5 to limit topic fragmentation across multiple clusters. The resulting clusters were considered *topics*. Next, Ifrim et al. introduced the term weighting scheme of Aiello et al. described above to rank the topics: *df-idf*$_t$ and the *entity_boost factor* are applied to the words in the clusters, relative to whether the words also featured frequently in previous time slots (Ifrim et al., 2014). Finally, each cluster was assigned a score, consisting of the highest term weight normalized by the cluster size, in order to be able to rank the clusters in each time window.

The events were then determined by selecting the earliest Tweet from each cluster among the top 20 for each time window, and applying another hierarchical clustering step to the resulting group of Tweets – but this time with a threshold of 1.0 to allow maximum cosine distance. This way, similar headlines got clustered together again, whereas former clusters that contained unique information remained unchanged (Ifrim et al., 2014). From the resulting clusters of headlines, examination of the respective keywords and the raw Tweets was used to validate whether the resulting clusters correspond to events reported in the media.

Clustering techniques have also been used to identify events by different types – such as recurring, known, and unknown events (Huang, Li, & Shan, 2018). An overview of clustering approaches in event detection with method, scoring, and system architectures can be found for further reference (Panagiotou, Katakis, & Gunopulos, 2016).

2.3 Clustering and classification

Contexts like social unrest can be identified accurately in Tweets using machine learning methods (Mishler, Wonus, Chambers, & Bloodgood, 2017). Therefore, a clustering step may be complemented with the application of a Naïve Bayes, Multi-Layer Perceptron, or the Decision Tree model to determine which clusters constitute an event that is relevant within a certain context (Walther & Kaisser, 2013). A general framework for combined applications (Panagiotou et al., 2016) is shown in figure 3.

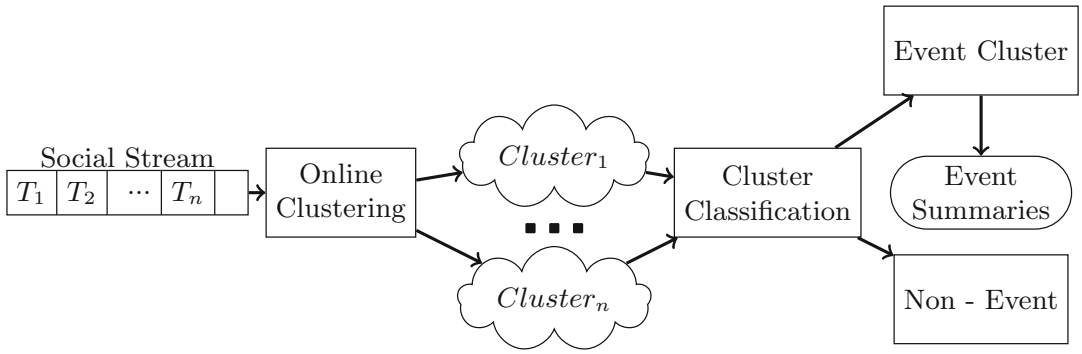


Figure 3: A general framework for clustering combined with classification; illustration from Panagiotou et al. (2016)

Others have changed the order, applying classification of Tweets first, then resolving the spatio-temporal aspects, and finally applying rule-based clustering for the identification of small-scale events. This approach has been used to discover fires, shootings, and vehicle accidents with great success: "Using 802k unfiltered tweets as provided by the Search API and the real-world data, we show that our small-scale incident detection approach detects more than 50% of the real-world incidents just using tweets. Furthermore, 32.14% of these are detected to be within a range of 500 m and 10 min to the real incident." (Schulz, Schmidt, & Strufe, 2015).

The approach taken by Schulz et al. can be considered conceptually advanced, as they used a probabilistic method enriched with external sources to accurately determine the location of the incident that is reported on the individual Tweet level – reportedly, at a median distance of 250 meters to where the event actually took place. Also, they applied a probabilistic method from previous research to determine the time of the event accurately from the Tweet texts. Furthermore, Schulz et al. were able to train their classifier on large collections of Tweets prelabeled as "car crash", "fire", "shooting", or "not incident related", which explains the good accuracy of the results. This approach seems to be highly suitable for the detection of previously unknown events by *known* categories.

Detailed work in the 2011 London Riots context has achiveved similar results for the detection of sub-events during the riot timeline: stabbings, shootings, store lootings, and protester gatherings in surrounding towns (Alsaedi, Burnap, & Rana, 2017).

For event detection by known categories, the TEDAS system (Li, Lei, Khadiwala, & Chang, 2012), illustrated in figure 4, can be considered a successful overall aproach: This system can accurately recognize criminal events like shootings, as well as disasters like tornadoes and floods. It first collects Tweets from the Twitter API by predefined keywords. The keyword vocabulary is then continuously expanded according to co-occurrence metrics of the keywords with other words and semantic links contained in the collected Tweets. Finally, a trained classifier is applied to detect whether a Tweet relates to one of the above categories of crime and disasters, displaying results on a map. This seems to be a highly effective approach for practical purposes, such as applications in law-enforcement and first response use cases.

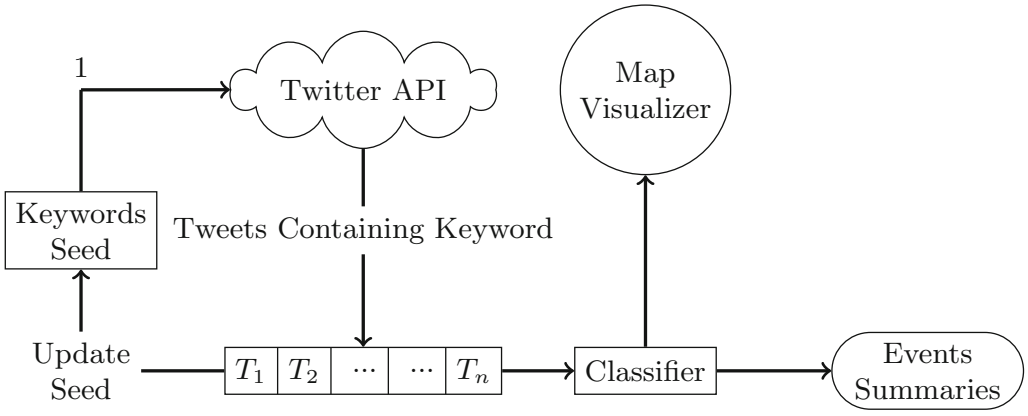


Figure 4: The TEDAS system; illustration from Panagiotou et al. (2016)

2.4 Advanced approaches

Korkmaz et al. have aggregated multiple data sources, including Tweets, news sources, the GDELT political database, Tor browser usage metrics, and currency exchange rates to a database, to which they applied a dimensionality reduction algorithm – the *Lasso* model – to predict civil unrest on a given day based on all the information available about a country from the respective previous day (Korkmaz et al., 2015). This approach was highly successful in predicting large scale social unrests, such as the protests known as the "Brazilian Spring", which took place in June 2013. Although their approach does not only rely on Twitter as a source, Twitter keywords were one of the top four sources chosen by the Lasso model.

In other study, *network analysis* techniques were used for earthquake detection and estimation of the corresponding epicenter based on the trajectory: Sakaki et al. use Support Vector Machines (SVM) for classification of the Tweets relevant to this context, and then apply Kalman and particle filtering to estimate the precise epicenter location based on Tweet-rates and timing (Sakaki et al., 2010).

Recall that Tweets can also be modeled as a graph or nodes and relationships between nodes (figure 5):

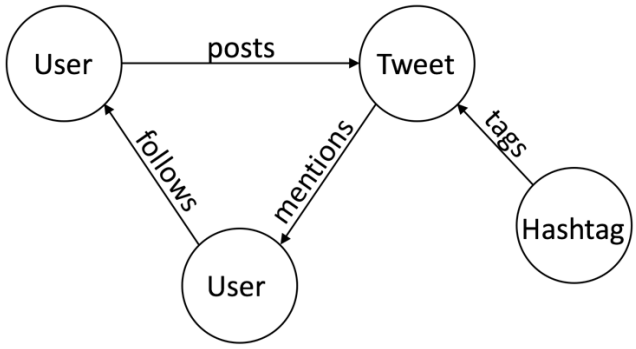


Figure 5: Tweets modeled as a graph; own illustration

Graph and network theory constitute a large field of research largely driven by the domains of physics and computer science. Many algorithms are available to support the tasks of detecting similar data within such a graph – from search approaches like PageRank (Page, Brin, Motwani, & Winograd, 1999) and RandomWalks (Wang & Landau, 2001) which use a centrality function to detect density, to community detection approaches like the Louvain algorithm (Que, Checconi, Petrini, & Gunnels, 2015) and the Label Propagation algorithm (Xie & Szymanski, 2011) which do not require a pre-defined objective function to accomplish this task.

Hua et al., have built an event detection system which first gathers vocabularies from labeled news reports, which distinguishes this approach as it allows the system to train a classification algorithm according to previously known event types, without the need for any manual Tweet labeling. They then applied several clustering, classification, and graph partitioning steps to identify events – complemented by a "wavelet method" used to measure geographic propagation of talk about the events to estimate event significance (Hua, Chen, Zhao, Lu, & Ramakrishnan, 2013). The wavelet concept is a mathematical approach from the field of signal processing in network analysis, which can be applied to a graph of Tweets. The underlying assumption is that the magnitude of an event can be measured by the reach of the information on the graph, i.e. the distance covered in nodes. The significance-estimation approach can therefore be considered similar to the above described earthquake epicenter estimation application of Sakaki et al., who essentially applied the same concept in reverse. The graph partitioning step used by Hua et al. was based on spectral graph theory, which constitutes a way of identifying clusters in a graph by considering both the evident relationships between

nodes and mathematically abstract concepts in the shape and patterns of a graph. Although the latter can be used as a heuristic, it can also be regarded as a bias: Becker argues that scalable graph partitioning algorithms perform poorly in isolating events from social media because these are sparsely distributed within the data, causing the algorithms to falsely assume a balanced shape (Becker, 2011). The Hua et al. paper does not include detailed information about performance or how the system capabilities were evaluated, which is perhaps explained by the fact that it was funded by the Intelligence Advanced Research Projects Activity (IARPA). However, the examples provided indicate that it was designed to be used as a querying tool for certain categories of events by region – eg. "civil unrest" in "Mexico" –which can be used on a database of historic Tweets or on a real-time stream of Tweets.

Weng et al. on the other hand identified events by isolating bursts of keywords based on their propagation pattern, which allowed them to cluster differing patterns of such signals on a graph. They then determined event significance by measuring the number of trending keywords in each cluster and calculating ther cross-correlation (Weng, Yao, Leonardi, & Lee, 2011). Based on this approach they built a system called "Voter's Voice" which was able to identify daily topics of discussion during the Singapore General Elections of 2011.

Finally, image analysis with deep learning has proven to be useful to classify street-level incidents like fire and physical violence based on images shared on social media in a riot context (Won, Steinert-Threlkeld, & Joo, 2017).

A note on use-cases and parties interested in the event detection topic:

Regarding the use-cases for the technology and solutions that have been developed so far, it is clear that governments agencies are interested in this topic. For example, several of the papers cited above were funded by the US government. Furthermore, material like the University of Maryland's *Handbook of Computational Approaches to Counterterrorism* includes descriptions of technical methods for forecasting political violence, including systems developed by the Defense Advanced Research Projects Agency (DARPA) (Subrahmanian, 2013). Similar material was presented by the Swedish Defense Agency at the 2011 European Intelligence and Security Informatics

Conference (EISIC) (Johansson et al., 2011). However, crisis and emergency response teams from different organisations around the world are using event-detection systems for situational awareness (Cameron, Power, Robinson, & Yin, 2012). In humanitarian applications, early-warning plays a major role for the efficient deployment of resources (Lopez et al., 2018) – which illustrates that information technology can be used for differing purposes.

2.5 Selection of approach

In order to select the most suitable method for the Gezi Park Uprising case, the methods were compared in a scoring table by order in which they were presented in the previous chapter. The suitability assessment was conducted based on an estimation of mathematical complexity, an indication of the transparency with which the individual steps were outlined in each respective paper, and whether the methods required labeled Tweets or an explicit geo-location feature (full reasoning in Table 1). For some points, the reasoning is included for extra clarity.

Paper	**Method**	**Mathem. complexity**	**Transpa-rency**	**Use of labeled Tweets**	**Use of geo-location**	**Suitability assessment for Gezi case**
Ranneries et al., 2016	Clustering	Low	High	No	Yes	Low; requires geo-location
Feng et al., 2015	Clustering	Medium	Medium	No	Yes	Low; real-time stream of Tweets used
Yang & Rayz, 2018	Clustering	Low	High	No	No	Low; number of events must be known in advance
Becker, 2011	Clustering	Medium, due to multi-dimensional approach	High	No	No	Medium, due to required learning effort and scope
Ifrim et al., 2014	Clustering	Low	High	No	No	Highly suitable
Huang et al., 2018	Clustering	High, due to complex clustering method	Medium	No	Yes	Low; requires explicit geo-location
Walther & Kaisser, 2013	Clustering & classification	Low	High	Yes	Yes	Low; requires previous knowledge of event types
Schulz, Schmidt & Strufe, 2015	Clustering & classification	High, due to complex location & time estimation	Medium	Yes	No	Low; requires previous knowledge of event types

Alsaedi, Burnap, & Rana, 2017	Clustering & classification	Low	Medium	Yes	Yes	Low; requires previous knowledge of event types
Li, Lei, Khadiwala, & Chang, 2012	Clustering & classification	Low	High	Yes	Yes, but could be applied without	Low; requires knowledge of event types
Korkmaz et al., 2015	Dimensionality reduction & regression	Medium	High	No	Yes, but could be applied without	Low; different event definition
Sakaki, Okazaki, & Matsuo, 2010	Classification & network analysis	High, due to complexity of epicenter estimation	High	Yes	Yes	Low; different use-case
Hua et al., 2013	Classification, graph partitioning & network analysis	High, due to combination of several complex methods	Low	No	Yes	Low; would exceed the scope of a thesis
Weng et al., 2011	Network analysis & graph partitioning	High, due to conceptual complexity of keyword burst identification	Medium	No	Yes, but could be applied without	Medium; could be attempted but the approach is very abstract
Won, Steinert-Threlkeld, & Joo, 2017	Classification with deep learning	Medium; due to parameter settings for neural networks	High	Yes	No	Medium; accessibility of Tweet images not clear

Table 1: Scoring table for the selection of the most suitable method

Based on this assessment, the hierarchical clustering approach described by Ifrim, Shi, & Brigadir (2014) was selected for the quantitative analysis of the Gezi Park data in chapter two. This method fits best to the task of detecting events of previously unknown category in an *unlabeled* set of Tweets within the scope of a Master thesis. Furthermore, the *df-idf*t_t term weighting scheme used by Ifrim et al. was considered the most practical approach for resolution of the temporal dimension.

3 Case: The Gezi Park Uprising

This chapter covers the quantitative part of this thesis, in which I applied the method described by Ifrim et al. to Tweets from the Gezi Park Uprising. This chapter begins with an introduction to the case to provide the reader with an overview of the uprising, followed by the step-by-step application of the method. Since the adapted CRISP-DM framework for data analytics was chosen as a structured approach for this data project, the subchapters after the case introduction follow the phases of this framework.

Istanbul's Gezi Park is a small green space on Taksim Square, which lies on the European side of the Bosphorus Strait. The Gezi Park Uprising began on May 28, 2013, following the announcement of the government that the trees in the park were to be cut down in favor of a new urban development project. As the bulldozers moved in at about 2 am local time, an assembly of 50 protesters gathered in the park in a peaceful sit-in to protect the trees. A few hours later, riot police evicted them from the park using what was described by the media as excessive force: participants were sprayed in the face with tear gas from a close distance and dispersed with water cannons to clear the way again for the bulldozers. Notably, MP Sirri Süreyya Önder was at the scene, accusing the police and the mayor of acting unconstitutionally in dispersing a democratic, peaceful protest. Following this incident, supporting demonstrations began with thousands of people taking to the streets in major Turkish cities to express their solidarity with the Gezi Park crowd. Within days, this wave of civil unrest turned into an anti-government movement, with many calling for Prime Minister Erdogan to resign. Istanbul was in an exceptional state during this time: Media reports described a festival-like atmosphere, and an unprecedented level of freedom of speech and social convergence. Comparisons were drawn to the Occupy Wall Street movement of 2011 and even the May 1968 events in France. On the other hand, violent clashes between supporters of the movement, riot police, and government supporters were reported on a daily basis. On June 10, PM Erdogan agreed to meet a delegation of protesters.

Figure 6 shows a timeline of Gezi Park Uprising events between May 28 and June 10, 2013. A map of Istanbul can be seen in figure 7.

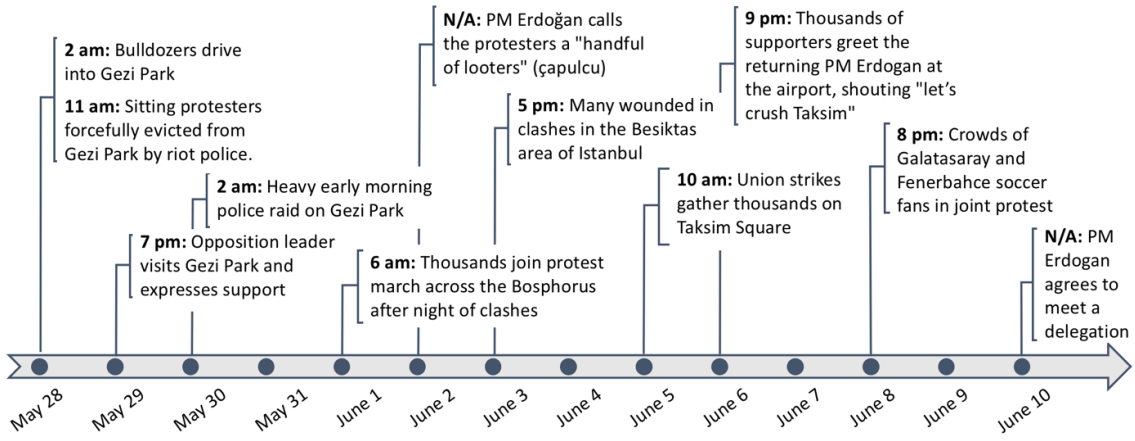


Figure 6: Timeline overview of Gezi Park Uprising events

(larger print version can be found in the appendix, A1)

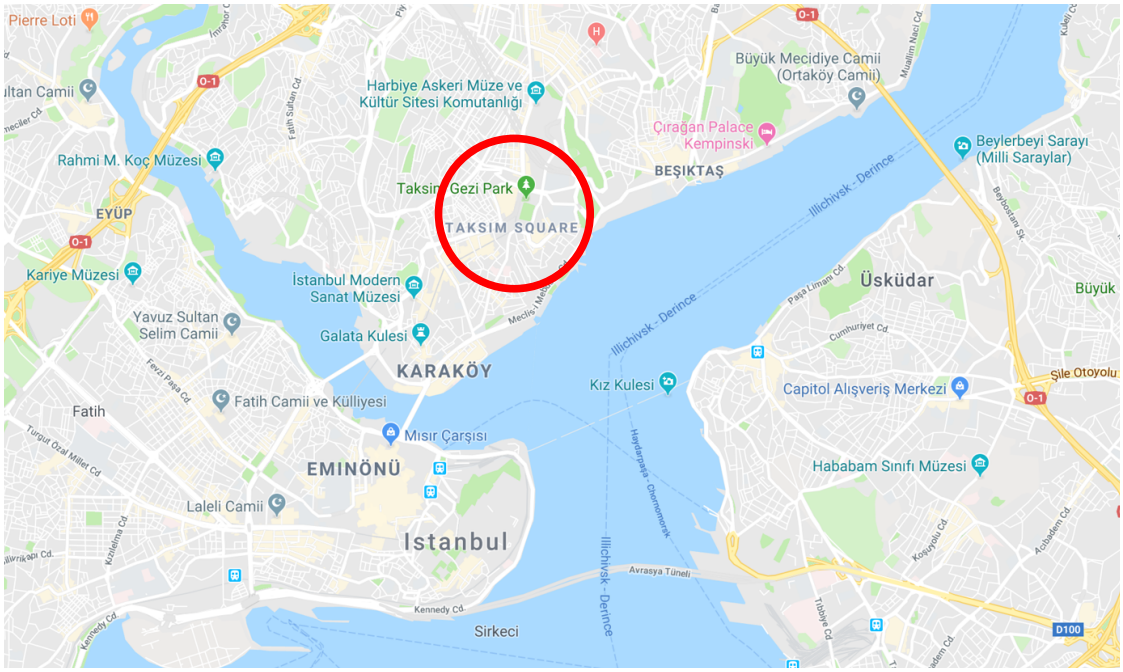


Figure 7: Google Maps view of downtown Istanbul; circle annotation indicates location of Gezi Park

3.1 Business understanding

In order to be able to apply the event clustering solution outlined by Ifrim et al. (2014) to the Gezi Park Tweets, a detailed understanding of the method is required beforehand. Once this is given, conclusions can be drawn regarding the data requirements, the technical system requirements, and regarding potential challenges that need to overcome in order to apply the method successfully. Furthermore, an evaluation scheme will be developed to measure the results – which are instrumental in answering the research question.

First, a more detailed survey of the method. The approach described by Ifrim et al. (2014) can be summarized in the following visual (figure 8).

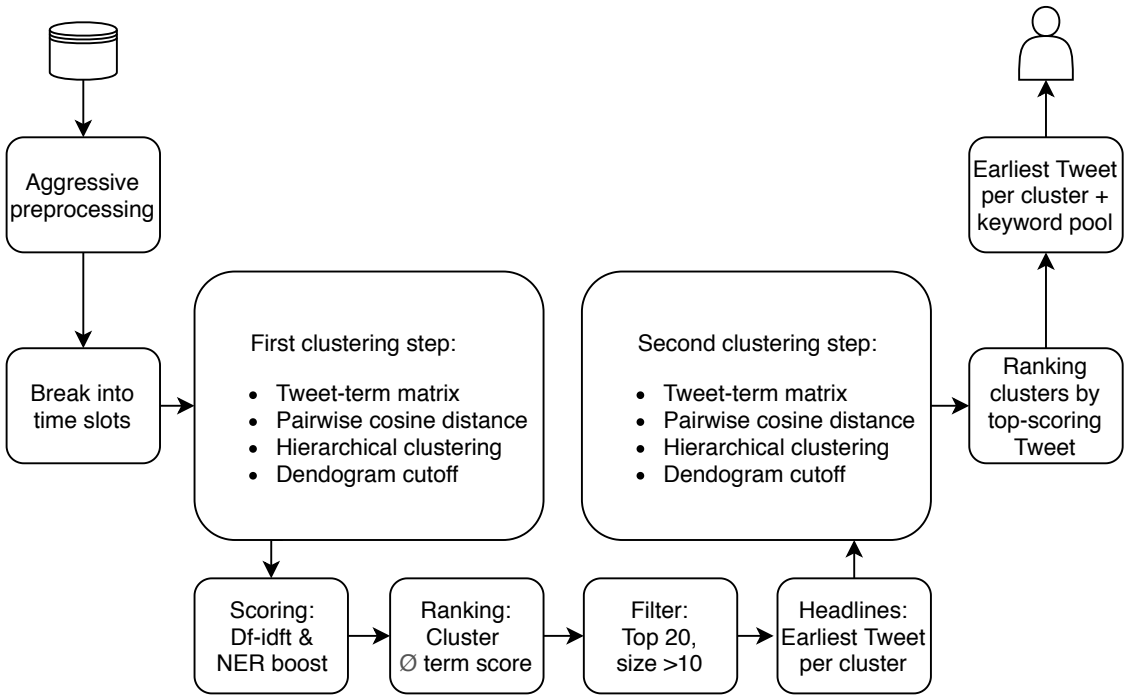


Figure 8: Visualization of the method used by Ifrim et al. (2014); own creation

Ifrim et al. applied this approach to the Tweets published in two different 24-hour timeframes during the days leading up to the 2012 US elections. Their goal was to detect election-related events (eg. “Romney wins Indiana”) in the data stream, as part of an organized data challenge in the context of this election.

As a first step, they "aggressively" preprocessed the data, including the removal of URLs, mentions, hashtags, digits, punctuation, whitespace, and the tokenization of the Tweet text. Furthermore, they removed all Tweet objects that had more than two mentions or hashtags, as well as all Tweets with less than four words. They then broke the datasets into 15-minute time slots – one example slot containing 9487 Tweets. For each time slot, they generated the Tweet-term matrix using the Sklearn library for vectorization, and calculated the pairwise distance arrays using cosine similarity as a metric. Recalling chapter 1.2.1, they only qualified terms that occurred in a minimum of *max(int(len(window corpus) * 0.0025), 10)* Tweets; meaning, an absolute minimum of ten which increases very slowly according to the size of the corpus. For example, if the window corpus has 10000 Tweets, a term would need to occur in at least 10000 * 0.0025 = 25 Tweets.

They then applied an agglomerative clustering algorithm from the *fastcluster* library to the resulting cosine similarity arrays, which yielded a clustering dendogram modeling the epochs of consecutive cluster merges. In order to produce the clusters, the dendogram must be cut at a certain height, which represents the average cosine similarity distance between document in the cluster – thus influencing the amount of clusters and their Tweet distribution. Ifrim et al. empirically determined 0.5 as the optimal distance threshold in order to achieve an optimal balance between the goals of having tightly defined clusters, and avoiding topic fragmentation across multiple clusters.

Next, the terms in each cluster were assigned weights according to the $df\text{-}idft_t$ method of Aiello et al. (2013), which penalizes the term frequency in the documents by the logarithm of the mean of its scores in the previous *t* time slots. Recalling chapter 1.2.1, a burst of similar keywords within a given time frame is ranked higher if the same keywords did not spike in the previous four time slots. Ifrim et al. went on to multiply the scores of all named entities by a factor of 2.5 instead of the 1.5 proposed by Aiello – with the reasoning that 2.5 produces more news-like topics. This way, trending non-event related topics (eg. recurring internet jokes) were automatically ranked lower (Ifrim et al., 2014).

The clusters were then ranked by their average term score and disqualified clusters with a corpus of less than ten Tweets. From the resulting list, only the top twenty clusters were kept, and from each of these clusters, the earliest Tweet was extracted as a "topic headline". This corpus of headlines was used for the second clustering step, which followed the same steps as the first one. However, this time the dendogram was cut off at the maximum distance to determine how many headlines were aggregated to a single topic.

These final clusters were simply ranked by the top scoring Tweet in the cluster, according to the initial term weights. In order to make sense of the output, Ifrim et al. extracted the earliest Tweet per cluster as a final "headline" from this ranking – displaying the raw text along with the respective cluster vocabulary as a topic summary. Using this approach, they found that 80% of the detected topics could be matched with news headlines during the respective period. They also stated that their tool took one hour to run on a 24 hour window, and that the main parameters influencing computing efficiency were the number of Tweets and the number of terms that were qualified in the respective steps, which both strongly affect the runtime for the cosine similarity and clustering computations (Ifrim et al., 2014).

The following is a high level analysis of the challenges that must be overcome in order to be able to apply this method to the Gezi Uprising case:

- **Data selection.** A Tweet sample needs to be acquired according to pre-defined selection criteria. The definition of criteria will be the first step in the *Data Understanding* phase.
- **Data accessibility.** Twitter has an API for developers which can be used to download Tweets. However, the API restricts the quantity of Tweets that can be accessed using this interface. Furthermore, this is associated with a cost of $ 0.99 USD per 100 Tweets, which would result in budget requirements of nearly $ 10 000 for 1 million Tweets. Since such a dataset was not found to be available online, two options remained: a) data acquisition from a third party, or b) using a scraping tool. The latter is a script which can search the Twitter "advanced search" page by applying the required filters, scrolling down, and saving all the search results in a JSON file.

- **Data cleaning and preprocessing.** Furthermore, Ifrim et al. (2014) mention "aggressive" preprocessing of data. The specific steps preprocessing steps required for the Gezi Uprising case will be determined based on the outcome of the *Data Understanding* phase.
- **Computing resources.** Natural Language Processing (NLP) can be computationally intensive due to the large amount of memory occupied by strings. In order to avoid long runtimes, efficient datatypes and code must be produced. Furthermore, the processing of a large dataset of Tweets can be expected to require powerful hardware – therefore, a scalable cloud server will be used as a computing environment. The cost of this *virtual machine* should be kept as low as possible due to funding constraints. Google Cloud Platform (GCP) was identified as the best solution, as this platform offers $ 300 in free trial credits which can be spent to deploy powerful and customized computing environments, with GPU and CPU resources available on a pay-as-you-go basis.
- **The Turkish language.** An existing approach which was devised to be used on an English dataset is to be reproduced with a Turkish dataset. This requires adaptation of the methods to the unique structures of the language. Both the availability of Turkish NLP tools, as well as the effort required to adapt existing tools or methods to the language are unclear at this stage – therefore, language presents a risk.
- **Transparency of the chosen approach.** The methods described in previous research may not be outlined clearly enough to allow simple reproduction of the technology (eg. copy and pasting code). For example, a research paper may introduce the mathematical concept and the logic of the algorithms used, without publishing the code. Therefore, it can be expected that a large portion of the work will be reproducing the required steps for preparing and processing the data in a way similar to the chosen approach from literature – which involves trial and error, testing functionalities, and choosing the right tools and solutions.

Regarding performance measurement and evaluation, Ifrim et al (2014) stated that they used a "subset of ground truth topics" as a basis for comparison, which had been provided by organizers of the data challenge in advance. They also compared the detected topics with news reports to determine relevance – finding that 80 % of the output from their system could be matched with media reports.

Due to the lack of a selection of "ground truth topics" to look for in the system output, a scoring approach was devised which enables both comparability to the work of Ifrim et al. and a reasonable level of objectivity. As described in the presentation of the research question, the success of the applied method will be evaluated by comparing the detected events to news reports about the events of the uprising, in order to determine whether a hypothetical user of the created event detection system would have been sufficiently informed regarding the security situation and the development of the uprising on a day to day basis.

As a definition for the term *event* the Topic Detection and Tracking (TDT) project definition was found to be suitable, as it connects events to consequences in the context in which they occurred:

> An event is "*something that happens at specific time and place with consequences*" (Allan, 2002).

In order to provide a basis for comparison, 64 news reports from the time period of analysis were assembled, in which events of varying levels of consequence were described. Using my own expert knowledge about the uprising and about the resonance which certain events had at the time, I distinguished the events by the categories of *major events* and *non-major events*. The corresponding tables of events provide a short summary description for each event, and the link to the news source in which it was reported (all reports retrieved in the week of March 25, 2019). Both tables are found in the appendix.

Major events were of consequence to the development of the overall uprising, while *non-major events* were judged to have been of consequence on other levels, i.e. the individual or the local level.

Using this list, I will analyze differences and similarities between events detected and events not detected to find out why this may have been the case. This will be complemented by an analysis of the rate of informative content returned by the system,

to provide a quantitative indication of performance regardless of the pre-selection of event reports (which can be considered biased).

3.2 Data understanding

The first step to understanding the data is data acquisition. In order to ensure the relevance of the data, a Tweet sample should be filtered from the total body of Tweets published during the relevant timeframe. For the objectives defined in this study, a relevant sample can be selected by checking for the use of hashtags relevant to the Gezi movement, and for the use of the Turkish language. This conclusion was reached based on the following assumptions regarding the data:

A1: Tweets from the time period between May 28, 2013, and June 10, 2013, which are written in Turkish and use hashtags connected to the Gezi Park Movement are more likely than other Tweets to contain information about the events, such as references to people and organizations, places, actions, objects, and descriptions.

A2: Tweets are an expression of the personal opinions and subjective perception of individuals and organizations, including supporters, opponents, as well as mere observers of the movement.

A3: Tweets contain *noise*, which can be characterized as non-informative and non-event-related content.

A4: The publishing time of the earliest Tweet within the subset of Tweets that are identified as similar event mentions can be used as a proxy for the time of the event. This approximation approach is reflected in the detection methodology.

A5: Location and origin of the Tweets: According to a previous study by New York University, about 90% of all the geolocated Tweets using the Gezi Park related hashtags originated from within Turkey, and 50% from locations in Istanbul (Social Media and Political Participation Lab, 2013). This is before language filtering, so it can be assumed here that a filter for *only Turkish Tweets* is sufficient to produce a geographically relevant data set.

In order to gain access to a large number of Tweets, option b) from *Business Understanding* was identified as the best solution: Download from the Twitter advanced search site via scraping tool. This was accomplished with a tool called Twitterscraper, which outperformed several other options found on Github. The Twitterscraper script takes four parameters as input: hashtag to search for, start date, end date, and the JSON output file name. It then downloads all the Tweets matching these criteria into the JSON file – in the exact order in which they appear in the advanced search results.

This method has two disadvantages by comparison to the API: firstly, only the features displayed on the regular Twitter frontend are accessible, and secondly, it is unclear if the advanced search results page displays the complete body of Tweets matching these criteria. The first can be seen as a constraint, the second is both a constraint and a limitation. However, the main difference to the API is the fact that the scraper only collects original Tweets: the information *how often* the Tweet was retweeted is supplied with each Tweet, but the corresponding Retweets themselves do not appear in the search as separate results. Conceptually, it is not clear whether this is an advantage or a disadvantage for our event detection approach. It could be argued that a retweeter is less likely to be an eye witness, eg. because an eye witness is too involved in the moment to read through other user's Tweets – but this would have to be verified empirically. In their analysis of the Gezi Park Tweet volumes, Varol et al. found that more than 50% of the overall Tweets published during the time frame of the protests were Retweets, and that the hourly volume of Retweets tightly matched the corresponding volume of original Tweets (Varol et al., 2014).

The start date of the protests was May 28, 2013. Although the protests continued for many weeks, the end date of the sample was set to June 10, 2013 in order to limit the volume of the dataset. Also, June 10 marks a turning point in the protests timeline as PM Erdogan announced he would meet a delegation of protesters.

The query hashtags were selected according to both news reports and research papers (eg. Varol et al., 2014) which analyzed the role of Twitter during the movement and stated the hashtags used. Using these parameters, I downloaded a total of 2247598 original Tweets. At this stage it should already be mentioned that this collection most

likely includes duplicates, occurring for each Tweet that has more than one of the collected hashtags. Table 1 reveals the number of Tweets acquired per hashtag.

Hashtag	Meaning	Count
DirenGeziParki	Rise up, Gezi Park	642 058
Gezi	Gezi (name of the park)	530 121
GeziParki	Gezi Park	348 731
OccupyGezi	Occupy Gezi	323 021
Capulcu	Looter	161 592
TayyipIstifa	Erdogan step down	92 094
DirenGezi	Rise up, Gezi	80 882
OccupyTaksim	Occupy Taksim Square	21 415
Istiklal	Istiklal Avenue	15 743
BugünTelevizyonlariKapat	Today, turn the TVs off	10 181
Capuling	Looting (play of words)	9 302
Yedirmeyecegiz	We're not buying it	7 463
Taksim	Taksim Square	4 016
Capulling	Looting (alternative spelling)	979
	Total	**2 247 598**

Table 2: Hashtags and corresponding number of downloaded Tweets

Figure 9 depicts the distribution of the Tweets over the time frame, grouped by one-hour time slots. As expected, the Tweet volume was low during the nights, following an even trend throughout the day, and increasing regularly after 6 pm local time (Coordinated Universal Time (UTC) +3).

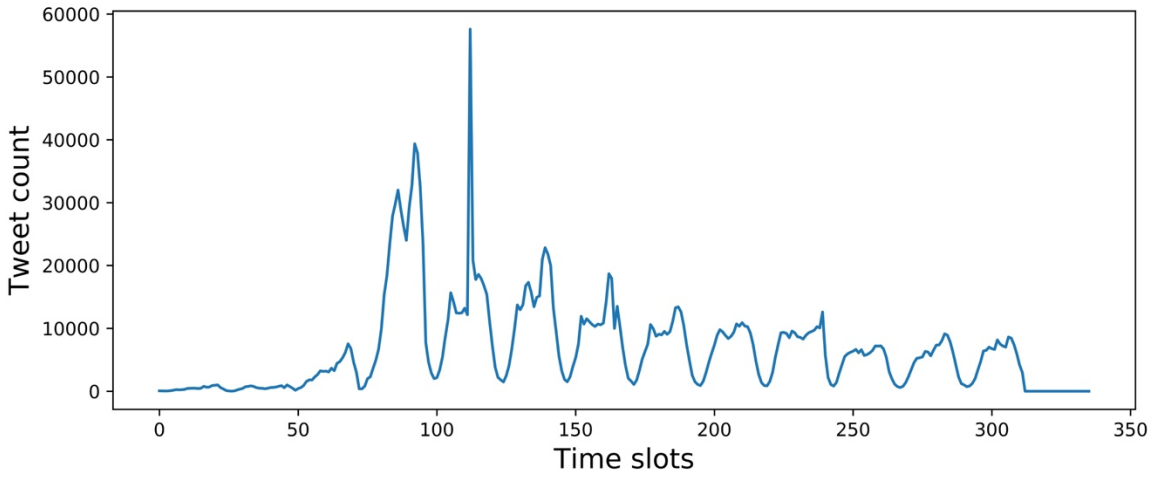


Figure 9: Hourly Tweet volume

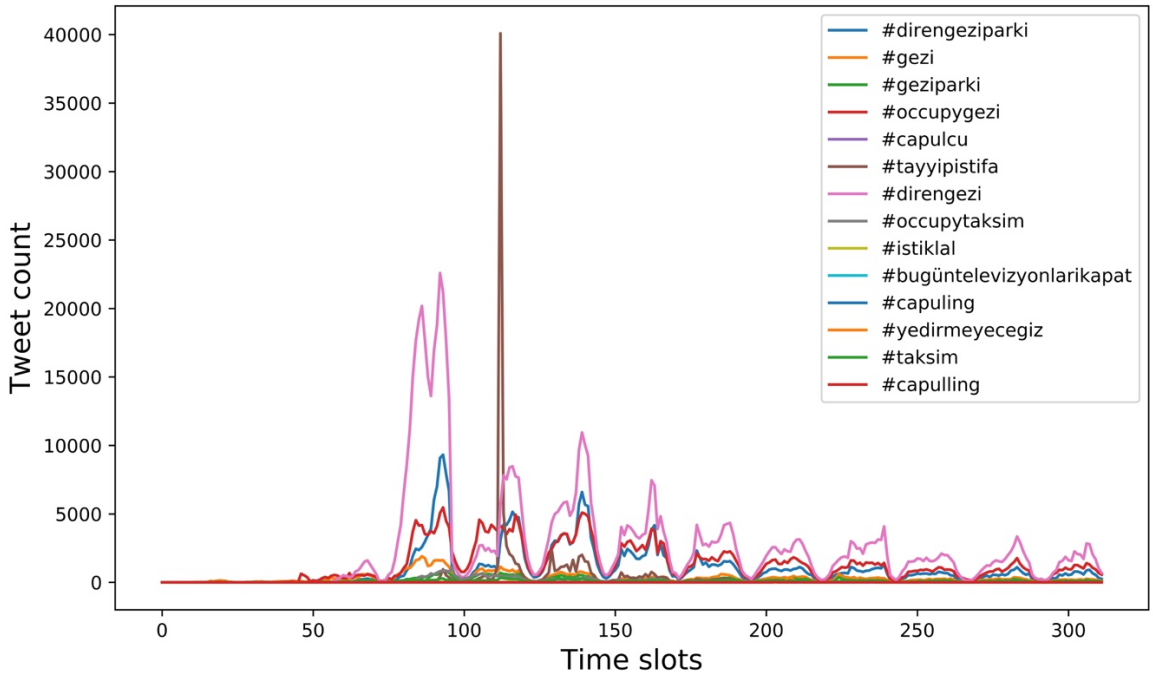


Figure 10: Hourly Tweet volume and hashtag distribution

Figure 10 shows the distribution by hashtags. Notably, most of the hashtags followed a similar curve – with the exception of #tayyipistifa, which translates to "Tayyip Erdogan, step down". This hashtag spiked considerably at around midnight in the night from June 1 to June 2, coinciding with intense clashes between protesters and riot police on Taksim Square.

Table 3 provides an overview of the data features collected, their data type, a content description, and the likely high level data cleaning steps required for each feature:

Feature	**Dtype**	**Content**	**Data cleaning steps required**
"fullname"	Unicode String	Full name	Drop to anonymize
"html"	Unicode String	Tweet as HTML	Drop – no use case
"id"	Unicode String	Numeric Tweet ID	Keep to identify duplicates and drop to anonymize after duplicates are eliminated
"likes"	Integer	Likes count	Drop – no use case
"replies"	Integer	Replies count	Drop – no use case

“retweets”	Integer	Retweets count	Keep, potentially relevant
“text”	Unicode String	The Tweet text	Hashtags, Emoji, links to web pages, and mentions of other users should be removed (substring). Further noise to remove includes special characters and misspelled words.
“timestamp”	Unicode String	Time published	Convert to timestamp
“url”	Unicode String	URL to the Tweet	Drop to anonymize
“username”	Unicode String	Username	Drop to anonymize

Table 3: Overview of data features, their properties, and required transformations

A first assessment of Tweet samples already revealed a number of problems, which need to be addressed in the *Data Preparation* phase:

- **Tweet language.** The dataset includes Tweets written in other languages, such as English and Spanish. Only Turkish Tweets should be considered to allow coherent term processing – therefore the language of the texts must be determined. If the text is written in a different language, the whole Tweet should be dropped.
- **Inconsistent spelling and character encoding.** The characters of the Turkish alphabet are encoded in the ISO-8859-3 standard, which is a subset of the Python-readable Unicode. Some of these characters, such as ş, ı, and ğ, are not part of the English alphabet. Although they are required for correctly spelled Turkish text, these characters are often substituted with their international counterpart in informal text redaction – possibly, because of keyboard size restrictions on a mobile device. Thus, “yoğun” (tough) for example may occur as the transliterated “yogun”. Without transliteration to a common standard, such variations cannot be treated computationally as the same word. Therefore, an additional preprocessing step is required to ensure coherent spelling.
- **Turkish grammar.** The structures of Turkish grammar uses suffixes to build sentences, resulting in longer and fewer words by comparison to English sentences. This has critical implications for stemming and lemmatization: For

English texts, many algorithms are available as the topic has been researched extensively (recalling chapter 1). In Turkish however, this is not the case. Furthermore, the suffixes can contain elements like verbs, pronouns, and negation – which means that information may be lost in suffix-stripping. This can be illustrated using the word *"gidebilecekler"*, which translates to "they will be able to go":

Element:	***"Gide -***	***- bile -***	***- cek -***	***- ler"***
Root:	"gitmek"	"bilmek"	"cek"	"ler"
Meaning:	"to go"	"to be able to"	future tense	third person plural

Table 4: Demonstration of the Turkish suffix problem

Through lemmatization, "gidebilecekler" could be reduced to "git", but important information would then be omitted. Therefore, this step needs to be handled carefully.

- **Spelling and capitalization.** In order to ensure the comparability of words used across a multitude of Tweets, spelling needs to be normalized. Necessary transformations may include lower casing and spell-checking.
- **Punctuation and conjunction.** Punctuation can be considered *noise*. In some cases, a score, dash, or apostrophe is used to join words together – thus, it convenes to replace all punctuation with a space to ensure the integrity of words. In the Turkish language, proper nouns frequently have suffixes attached to them with an apostrophe, for example: "Gezi Parkı'ndan", which translates to "from the Gezi Park". Removing the apostrophe makes it easier to recognize the object *Gezi Park*, and *ndan* can be treated separately.
- **Concurrence of names, places, and things.** Turkish given names frequently concur with names of places and things, because they carry literal meaning: For example, "yunus" translates to *dolphin*, but "Yunus" is also a very common given name. It could be a reference to the member of parliament *Yunus Emre*, and "Yunus" is also the name of a light rail station in Istanbul. This has implications for any attempt to extract mentions of people or locations from the Tweet text.

- **Non-word elements.** Tweets include mentions, hashtags, web links, emoji and special characters. These items should be removed as they are ambiguous and cannot be interpreted clearly within the scope of this thesis.
- **Stopwords.** Many words do not contain information and should therefore be removed, for example the Turkish counterparts for “and”, “so” or “this”. That can be achieved by running a list of such words against the tweet texts. The stopwords list should also include the endings that are separated from the words through replacement of punctuation with empty space, as mentioned above. It can be assumed that the informative part is contained in the nouns, verbs, adjectives, and pronouns, because these contain the answer to the *who*, *what* and *how* questions that explain an incident. *Why* is an out-of-scope aspect which is difficult to extract reliably, and *when* shall be answered using the timestamp as a proxy (as outlined in assumption A4). In order to minimize noise, the content for analysis should be reduced to the four categories of words described above.

3.3 Data preparation

As a first step, the work environment is prepared. For data projects in Python, the Anaconda package manager offers a convenient solution to build programming environments with many of the required packages pre-installed. First data processing steps were undertaken on a MacBook Pro with a 2.4 GHz Intel Core i5 processor and 4 GB ram, which was too slow. Therefore, this setup was quickly replaced by a scalable Google Compute Engine environment built to the following specifications (Table 5):

Component	**Chosen Specification**
Datacenter Region	europe-west-1b (Belgium)
Machine Type	n1-standard-96
Operating System	Linux Ubuntu 16 TLS
CPU Platform	Intel Skylake
CPU Count	8 – 96 (variable)
CPU Memory	120 - 360 GB
GPU Count	2 – 4 (variable)
GPU Type	Tesla T80
Boot disk size	250 GB

Table 5: System environment used

Following the set-up of the Anaconda environment and the required standard modules, the data was migrated to the server. Jupyter Notebooks were used as an interactive shell environment for the data exploration and manipulation with Python, and the Numpy and Pandas modules. An overview of software packages used can be found found in table 6.

Software Package	**Functionality**
Anaconda	Package manager for scientific computing in Python and R
Python 3.7	Programming environment
Jupyter Notebooks	Open source server application for interactive computations
Pip	Package finder and dependency manager for Python libraries
cURL	Software library command interface
Git	Version control tool and Github interface
Homebrew	Package manager for Linux
Icu4c	C/C++ and Java libraries for Unicode
PyICU	Python extension wrapping the ICU C++ libraries
Pycld2	Python bindings around Google Chromium's embedded compact language detection library
Morfessor	Unsupervised morphological segmentation tool
NLP Cube	Neural network-based Adobe NLP tool
Cython	Static compiler for Python and Cython
Numpy	Program library for scientific array computations
Pandas	Program library for structured data processing
Tweet-Preprocessor	Package that can remove specific content from Tweet text bodies
Polyglot	NLP tool that can detect named entities in Turkish text
Langdetect	Neural network-based language detection module of Google Translate
Multiprocess	Parallel processing framework

Table 6: Software packages installed

Based on the problems identified in the previous subchapter, a preprocessing pipeline was designed to execute all the necessary steps to produce a dataset suitable for modeling according to the Ifrim et al. approach. Recalling their paper, the first step was “aggressive” preprocessing and filtering. In my case, this was not only necessary because of the modeling approach, but also in order to deal with the language-specific problems identified in the previous subchapter. Thus, the pipeline had to fulfill the following tasks:

Task	Dependencies	Possible solution
Removing duplicates	Should be done early, for efficiency	Python
Making “timestamp” feature readable by Python	None	Python
Removing mentions, hashtags, links and emoji	None	Python, or available tools; to be determined
Removing all Tweets written in languages other than Turkish	Should be done early, for efficiency and accuracy	Available tools; to be determined
Removing all punctuation from the “text” feature	After language detection, to avoid impairment	Python
Extracting named entities from “text” into a new feature (NER)	After punctuation removal; before text standardization and normalization steps	Available tools; to be determined
Removing stopwords from “text”	After NER	Python
Extracting location mentions from “text” into a new feature	After stopwords; before location and lemmatizing/stemming - efficiency gain	Available tools; to be determined
Extracting all verbs, nouns, adjectives and pronouns from “text” into a new feature	After stopwords, location mentions and NER	Available tools; to be determined
Normalizing all the new features by tokenizing, transliterating, and lower-casing the words	In the end, or within functions of other points to ensure differently spelled versions are all considered the same	Python, or available tools; to be determined
Standardizing the words by reducing them to their stem or lemma, without creating too much noise or losing too much information	At extraction point of relevant words (verbs, nouns, adjective, pronouns)	Available tools; to be determined
Aggregating the outputted features to one word-list per Tweet object, including named entities, location mentions, and standardized words	In the end, after all other steps	Python

Table 7: Required preprocessing steps

Based on this table, the following process pipeline was found to lead to the desired output:

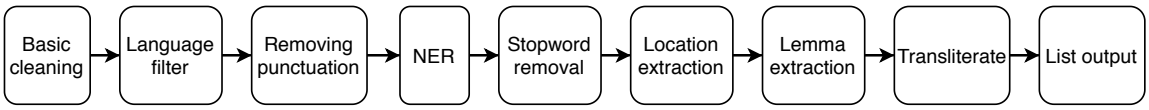


Figure 11: Preprocessing pipeline; own illustration

In the next sections, each of the steps is described shortly to explain the design choices – and where applicable, the respective software solutions used and how these were tested.

Basic cleaning. This step includes the removal of duplicates, the conversion of the timestamp from a string feature to a Unix timestamp, dropping all unwanted features (including personal data), and the removal of links, emoji, mentions, and hashtags from the text feature. The first three were accomplished using Python syntax and Pandas, whereas the fourth part was achieved using a tool called Tweet Preprocessor – a Tweet cleaner found via Github which performed well on test samples. This step was placed at the beginning to make all the following steps more efficient. As expected, the portion of duplicates was significant, covering 23% of the dataset:

Duplicate Tweets dropped:	507867
Remaining Tweet count:	1739731

Table 8: Changes to the dataset due to basic cleaning

Language filter. The goal of this step is to detect the languages used in each Tweet and to filter for only Turkish Tweets. This should happen early in the pipeline, as following modules (eg. NER and lemmatization) are language-specific operations. Langdetect is a Python implementation for the language detection module of the Google Translate service, which is originally written in Java. The Langdetect code is available on Github (Michal Danilák). Testing runs with 1000 Tweets yielded satisfactory results with very few false classifications: failure to accurately detect the language appears to be closely

linked to multiple spelling errors, unusually short Tweets, or the usage of multiple languages in a single Tweet. Since it can be assumed that such cases are less likely to contain valuable information, this problem can be disregarded – it simply led to a further reduction of the overall Tweet sample:

Number of Tweets dropped:	459873
Remaining Tweet count:	1279858

Table 9: Changes to the dataset due to language filtering

Removing punctuation. A Python function was used to replace all punctuation marks with a space, in order to ensure the separation of words.

Named Entity Recognition (NER). NER refers to the recognition and tagging of entities mentioned in text. These mentions of people and organizations can be considered informative content regarding events, and should therefore be extracted early in the pipeline to prevent loss of the exact name through suffix removal. A wide range of tools are available to accomplish NER with deep learning methods, but only very few of these are trained on a labeled body of contemporary Turkish text. The tested tools are presented in table y with an assessment of tests result on a sample of Tweets:

Tool	**Creator**	**Assessment**
ITU NLP web-service	Turkish NLP tool developed by the Istanbul Technical University (ITU)	Satisfactory performance in web-app, but the API is not open – which makes batch processing impossible.
Polyglot	Multilingual NLP toolkit developed by developer Rami Al-Rfou	Accurate performance with only few exceptions related to lower-cased names. Reliably detects *persons*, *organizations*, and *places*.
Turkish POS Tagger	Turkish Part of Speech tagging tool developed by Onur Yilmaz	Reliably labels each word by category (including proper nouns); was found to be difficult to deploy in a Jupyter Notebook.
NLP Cube	NLP suite of tools developed by Adobe	Can detect proper nouns in Turkish text with a reasonably high degree of accuracy (Boroş, Dumitrescu, & Burtica, 2018).

		However, it is computationally expensive as it considers the context of each word.
	Selected tool:	Polyglot

Table 10: Decision table for the NER tool used

Stopwords. Words such as “and”, “this”, “because”, and “whenever” do not contain information in themselves without consideration of context. These words should be removed at this stage to allow for greater efficiency in the following steps. In order to achieve this, several lists of Turkish stopwords which were created and published by researchers and developers were aggregated and then reduced to contain only a set of unique stopwords. This list was further enriched with the Turkish suffixes that were separated from their words following the punctuation-removal step, as well as other additional stopwords that I discovered during test runs. Using list processing and a function that iterates though each word of the Tweets, stopwords were successfully removed.

Location extraction. Since the metadata of the collected Tweets did not include the geolocation feature provided by the Twitter API, the best proxy for the event location was found in the location references in the Tweet text. This approach gives credit to the fact that people may disseminate information about locations beyond their immediate surroundings. However, not all locations can or should be considered – therefore, lists were created of locations that may be relevant for this business objective. In this case, Turkish cities were already identified accurately by the NER module, so the high level view was already covered. In Istanbul, where most of the events took place and the available media coverage was more detailed, the granularity of event analysis should be more precise to include any possible references to places that present meaningful points of interest to the locals. Therefore, a list of the names of all neighborhoods, main streets, squares, public transport stations, parks, bridges, harbors, mosques, and other important landmarks was produced manually, based on a number of web sources and personal knowledge of the city. For the extraction, a reverse matching method was applied: For each word in the list, if a match was found in the Tweet token strings, the word contained in the list was appended to a new feature containing detected location

mentions. This was found to be the best way to extract the locations without losing the mentions which had suffixes attached to them.

Lemma extraction. Lemmatization is a text normalization technique used in NLP. The goal of this step is to make word occurrences as comparable as possible, by reducing them to their lemma, or stem. The difference is subtle: a lemma has to be the actual root word, whereas the stem is simply the result of removing suffixes – sometimes not resulting in a proper word. For example, the stem of the word "troubling" may be "troubl", after the removal of the suffix "–ing". The lemma, or root word, in this case is "trouble". As mentioned in *Data Understanding*, achieving this in Turkish presents a challenging task as many words use multiple suffixes, which results in a risk of losing too much or too little information – depending on which elements are recognized as a suffix. The following options were surveyed and tested:

Tool	**Creator**	**Assessment**
Turkish-Lemmatizer	Tool developed by Abdullatif Köksal based on the Zargan Dictionary (Bilgin, 2016), a lexical database containing 1.3 million Turkish words with stems	Insufficient, but usable due to output of several options (examples further below)
TurkishStemmer	Turkish stemming tool developed by Osman Tuncelli	Insufficient
NLP Cube	NLP suite of tools developed by Adobe	Can lemmatize correctly spelled Turkish words with an accuracy of 87.84% according to the documentation (Boroș et al., 2018).
	Selected tool:	NLP Cube

Table 11: Decision table for the lemmatizer used

The table below contains a sample of test results using the Turkish-Lemmatizer tool, which returns several output options:

Input	**Translation**	**Output 1**	**Output 2**	**Output 3**
göremedik	we couldn't see	*göreme*	*göre*	*gör*
kolaylaştırmak	to make … easy	*kolayla*	*kolay*	*kol*
satıcıları	the vendors	*satıcı*	*satı*	*sat*

Table 12: Turkish-lemmatizer testing results (excerpt)

In the first and third examples, Output 3 can be considered an accurate stem: “gör” translates to “see”, and “sat” to “sell”. In the second example, Output 2 would be more useful, because “kolay” or “easy” can be considered the lemma of the word. Output 3 in this case is problematic, because “kol” coincidentally translates to “arm”, therefore leading to a false interpretation. Although a satisfactory algorithmic solution to choosing the right output could be achieved for the Turkish-Lemmatizer, the NLP Cube outperformed it in test runs on 1000 Tweets. Another reason why this tool was considered superior is the fact that it offers additional functionality, such as the tagging of the word category and inclusion of attributes relating to the function of the word in a sentence.

Table 13 contains the raw Cube outputs for the following sample Tweet text:

> *“Bu defa dik duracağız, korkmayacağız hiç kimseye yedirmeyeceğiz.”*
> (“This time, we will stand upright without fear and we will not have anyone ‘buying’ this” – emphasis added.)

Index	Word	Lemma	XPOS	UPOS	Attributes
1	Bu	bu	DET	Det	_
2	defa	defa	NOUN	Noun	Case=Nom|Number=Sing|Person=3
3	dik	dik	ADJ	Adj	_
4	duracağız	dur	VERB	Verb	Aspect=Perf|Mood=Ind|Number=Plur|Person=1|Polarity=Pos|Tense=Fut
5	,	,	PUNCT	Punc	_
6	korkmayacağız	ka	VERB	Verb	Case=Gen|Number=Plur|Person=3|PronType=Dem
7	hiç	hiç	ADV	Adverb	Case=Nom|Number=Sing|Person=3
8	kimseye	kimse	NOUN	Noun	Case=Dat|Number=Sing|Person=3
9	yedirmeyeceğiz	yedir	VERB	Verb	Case=Nom|Number=Sing|Number[psor]=Sing|NumType=Ord|Person=3|Person[psor]=3
10	.	.	PUNCT	Punc	_

Table 13: NLP Cube testing results (excerpt)

The lemma module had satisfactory performance on all word types except verbs, for which the Cube seems to output the word stem by definition, not the lemma. However, this does not influence the comparability of words and can therefore be ignored. In the above example, only the word in index six was not processed correctly (the stem should have been "kork"). The XPOS feature has a good score at 90.17%, according to the documentation. As a final solution implemented for this pipeline step, the lemmas of only nouns, verbs, adjectives, and pronouns were extracted by matching the XPOS field to these word categories. Thus, the noise in the output feature was further reduced.

A note on computing resources:

NLP Cube uses pre-trained neural networks for the respective tasks of classifying the word category and determining the lemma. Initially, running this pipeline step on each of the words contained in the Tweets would have taken more than 100 hours. Through parallelization of the compute job onto 96 CPU cores, runtime was effectively reduced to 6 hours and 27 minutes for the whole dataset.

Transliteration. In this step, the text in the output feature of the NER, location extraction, and lemma extraction steps was normalized by replacing all the Turkish characters with their international counterpart. This was effectively accomplished with a Python function that iterates through each character in the Tweets, and a dictionary containing the 'key, value' pairs for character mapping (eg. 'ı, i'). Applying this step in the end ensured that all different spelling versions of words were considered the same word – with the exception of spelling errors, which still produced noise (compare evaluation after the modeling phase).

List output. At this point, the outputs from NER, location detection, and lemma extraction were aggregated to a list and stored in a new column containing all the results. Duplicates were removed for each Tweet; therefore, the only noise left in these lists can therefore be described as a) word types that were falsely classified by NLP Cube, b) lemmatization failures, and c) some extreme cases of falsely extracted locations. *Category a* affects about 20 % of words – potentially more, considering that spelling errors have a negative impact on the NLP Cube's performance. However, false classification to another target category (eg noun extracted as VERB) does not present a

problem, as the word would have been extracted anyway, resulting in the same outcome. Therefore, the overall effect of this category is estimated to be below 20 %. *Category b* – false lemmas – affects 12-13% of cases according to the documentation. This results in a larger sample of total extracted words due to differing lemma variations, which certainly had a negative impact on the overall results of the study (compare evaluation). However, solving the issue would require a highly complex operation; therefore, this noise had to be accepted as an accuracy constraint. *Category c* is a collection of rare location-related misunderstandings. For example, a bus stop named "Emek" ("pension") was extracted falsely every time the verb "demek" ("to say") was used. Therefore, Emek was simply removed from the locations list – while it is safe to assume that other misunderstandings of this kind persist.

The output from the Data Preparation phase was a table containing the index, raw text, timestamp, NER, and the combined tokens from the last step for each Tweet.

3.4 Data Modeling

With the preprocessed data set in place, the data modeling approach of Ifrim et al. (2014) can be applied to the Gezi Park data set. However, the following key differences were identified beforehand, requiring adaptations of the approach:

- The Gezi Park dataset covers a time span of thirteen days, as opposed to the twenty-four hours processed by Ifrim et al. Apart from the impact on computing resources, this difference had conceptual implications for the applicability of the strategy outlined for the second clustering step, in which top-ranked clusters of the time slots were collated into final "topics". This does not make sense for a 13 day time period, therefore the re-clustering step was applied on a day by day basis to match the time span of 24 hours, analog to Ifrim et al.
- Election-related events follow a different event definition than the one chosen for my case: they happen at a highly specific time, eg. considering the breaking news announcement of an election winner for a given state, but not necessarily at a specific place. In the Gezi Park case, the locations of many of the reported events were highly specific. Therefore, location mentions had to be taken into account – which was achieved with the location extraction step in the preprocessing phase.

In order to accommodate the event definition chosen for my approach, I first experimented with time slots of two hours. Two hours seemed a good minimum, keeping in mind the goal of reducing topic fragmentation across time slot borders, which may cause problems for events that went on for longer than a few hours: Such fragmentation would lead to a different cluster in each time slot for the same event, with the cluster in the second slot ranked lower due to recurrence of the topic (as a result of $df\text{-}idf_t$). This did not seem conceptually problematic for the detection of the beginning of an event, but the positioning of the corresponding cluster in the ranking of the consecutive slot was not as clear – potentially limiting the ability of the method to detect such events. Therefore I initially tried to apply the steps to two-hour slots.

While these could be computed without problems for most time slots – during which the overall number of Tweets collected ranged between 50 and 5000 Tweets – the

processing power and time required for the larger arrays found in certain time frames were too high, especially for the cosine distance and the first clustering step. This is reasonable, as the number of calculations required to assess the similarity of Tweets grows exponentially with the increase in number of Tweets and qualified terms. For example in the first clustering step, I found that the clustering took less than 0.05 seconds for slots with up to 2500 Tweets, about 70 seconds for 4500 Tweets, about 5 hours for 20000 Tweets – and for corpuses in excess of 25000 the feasibility could initially not be determined. Since the largest corpus had 49408 Tweets, processing two hour slots was not possible: even distribution across 96 CPU cores did not produce a result in reasonable time. A reduction of the number of input terms may help improve performance, but this would need to be investigated further.

Figure 12 shows the variance of Tweets resulting from cutting the data into two-hour slots:

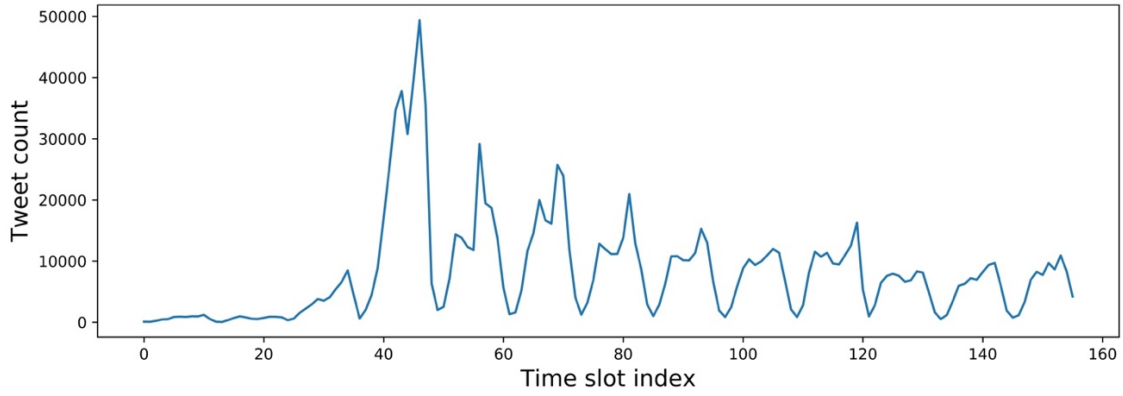


Figure 12: Tweet volumes by 2-hour time slots

For these reasons, I switched to one hour slots, which on average cut the amount of Tweets per cluster in a half, yielding 312 slots of between 14 and 25493 Tweets. This time it worked, but runtime was still long at about 34 hours for the pairwise cosine similarity calculation of the documents and the first clustering step. To illustrate the amount of calculations that this involved: The cosine similarity step yielded 41 gigabytes of numeric arrays, even though *float16* encoding was used instead of the recommended *float64* data type. Based on my experience with test runs I estimated that over 95 % of the final runtime could be attributed to the processing of just four

consecutive hours between 6 and 10 pm UTC on May 30, 2013 – which each had more than 20000 Tweets.

However, variance also proved to be an obstacle for the smaller time slots, because for some slots no terms were extracted due to the requirement of occurrence in a minimum of *max(int(len(window corpus) * 0.0025), 10)* Tweets. I overcame this problem by relaxing the absolute minimum to eight Tweets.

The next challenge was determining the right height at which to cut the dendograms. A 0.5 distance threshold as proposed by Ifrim et al. was not found to work well, leading to only a single cluster in most slots. Figure 13 contains the visualization of such a dendogram, showing the epochs of agglomerative clustering in a smaller time slot. The height (y-axis) here indicates the average cosine similarity value between all the document pairs in a cluster, whereas the sample indices (x-axis leaflets) denote individual Tweets. The junctions in the graph indicate merges between clusters – so the graph should be read starting from the leaflets on the bottom. The jagged horizontal line annotation represents an example cutoff height of 0.055, which would result in two clusters: cluster *a* and cluster *b*. This height can be raised or lowered in order to determine the best fit in the trade-off between creating a small number of large clusters with fragmented topics, or achieving a high number of small, keyword-focussed clusters. In this example, slightly raising the bar to about 0.065 would already mean that only one cluster would be returned, including *all* Tweets in the time slot.

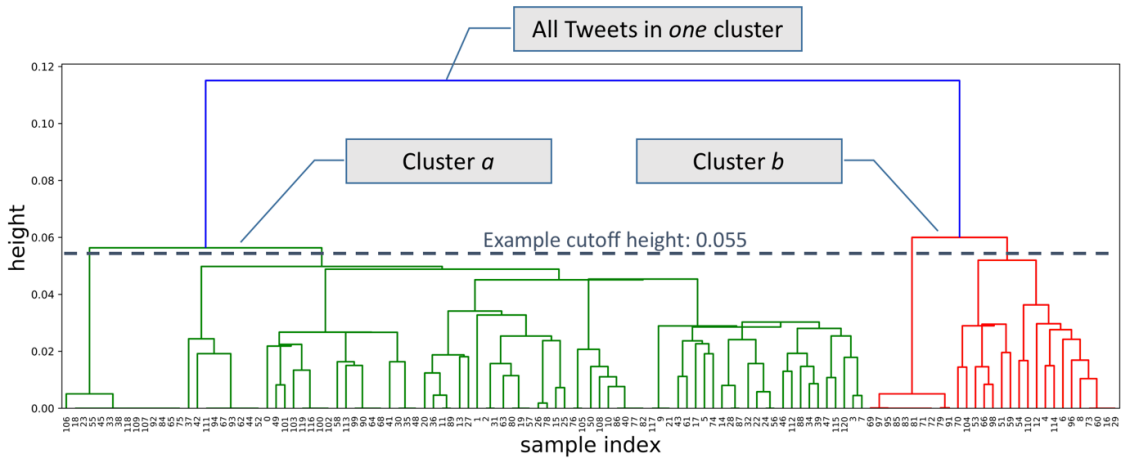


Figure 13: Example dendogram, annotations added

Due to the large variance in Tweet volume in the slots, it was impossible to find a single value that worked well for all sizes. As defined by Ifrim et al., clusters should have a minimum of 10 Tweets to constitute a topic. In order to attain a reasonable balance between cluster size and the amount of clusters returned for each time slot, a customized cutoff height was introduced, which used a *divident* to fit the cutoff height to the count of Tweets in each time slot. Table 14 depicts example cutoff heights which were found to be optimal, leading to a consistent set of clusters.

Slot size	Cutoff height	Cluster count, n > 10
700	0.005	5
1800	0.002	15
5000	0.00075	25
25000	0.00015	110

Table 14: Determining the optimal cutoff height

The relationship between the slot size and the cutoff height was in this case described by the function y = 3.5/x, where y is the cutoff height and x the slot size. Thus, applying a dividend of 3.5 was found to work best: 3.5/700 equals 0.005, and 3.5/1800 equals 0.002, which respectively yielded 5 and 15 qualified clusters. The cluster count increase was found to be proportionate to the increase of slot size with this estimation method, yielding between 2 and 111 clusters with a minimum size of 10 Tweets – depending on the amount of Tweets published in the hours and the respective data. In total, 5701 qualified clusters emerged; an average of 18 clusters per time slot.

For the calculation of the term weights with $df\text{-}idf_t$ I also used the previous four slots as a basis for comparison. For the first four slots all the *existing* previous slots are are used, so for the first slot the weighting is based on simple document frequency.

Having applied weights to the slot vocabularies, a first look at the wordclouds of the slot vocabularies revealed that the important terms indeed reflected words which could be linked (figure 14): For example in time slot 85, which is 3 pm local time on May 3, prominent terms in the vocabulary include gezi, park, taksim, “toma” (Turkish abbreviation for armored police vehicle with water cannons), water (“su”), hit/blow

(“darbe”), person (“insan”), everyone (“herkez”), and gas (“gaz”). Furthermore, there seems to be something about a lawyer (“avukat”) and about the bar association (“baro”).

Figure 15 pictures Taksim Square at the same time, showing armored police vehicles and clouds of tear gas on Taksim Square, indicating that this vocabulary can be directly linked to events that took place there.

Figure 14: Slot 85 word cloud

Figure 15: Taksim Square pictured on May 31, 2013; photo courtesy of Cem Yilmaz

Paylaşınki Herkez Görsün ! Toma Darbesiyle Yaşamını Yitirdi ! Katledildi ! youtube.com/watch?v=_uofQz... #direngeziparkı #dayangeziparkı

Figure 16: A sample Tweet from time slot 85, suggesting that a man was forcefully hit by a water cannon.

News reports indicate that many were injured in violent clashes between protesters and police on Taksim Square in the afternoon (Turkish Weekly, France24), but do not mention specifically that a man was forcefully hit. As a next step, sample clusters from this time slot were be analyzed to verify that the clusters can be described as topics, which was suggested by Ifrim et al. A number of random samples were surveyed, confirming that this is the case. For the specific example shown above, a cluster was found which precisely isolated the keywords describing the specific information contained in the sample Tweet (figure 17), indicating that at least nine other Tweets also described such an incident. However, many of the clusters were not as informative, containing random keywords connected to the uprising (such as figure 18), or they could be matched with other events on May 31.

Figure 17: Informative cluster

Figure 18: Non-informative cluster

Next, the clusters were ranked by average term weight and the earliest Tweet in each top-20 cluster was selected as a topic headline, resulting in a table of cluster representatives. Ifrim et al. simply went on to run the clustering algorithm on this output. However, to avoid joining together topics out of different days, the table of cluster representatives was split up according to the days, allowing separate clustering of the headlines from each day. Several cutoff heights were tested to optimize the trade-off between a reasonable amount of clusters and topic precision, but in the end the same method as in clustering step one was found to work well – applying a dividend of 3.5. This yielded 185 clusters, which I considered a reasonable number for an analysis of events over a thirteen-day-period.

Analog to Ifrim et al., the earliest Tweet from each final cluster was used as a headline. However, this method led to considerable gaps in the distribution of the headlines across the hours of some of the respective days. Therefore, I decided to additionaly include the Tweet with the highest term weight score. This on one hand added a comparatively informative Tweet from each cluster to the selection in the final table, and on the other hand led to a better coverage of the different times of day. The fact that using only the first Tweet from each cluster would have produced long gaps in the span of some of the days is an indication that on these days, the events spanned over multiple hours as different slots had clusters of similar keywords.

Additionaly, I did not only use the vocabulary list from each final cluster for interpretation of the results, but enhanced the Ifrim et al. method by retrieving the term weights according to the respective time slot in which the selected headline Tweets were published, to generate a weighted vocabulary for each final cluster. This allowed a novel, combined way of visually reviewing the final headline cluster vocabulary through the lens of how important the respective words were in the specific time slot of the *headline* Tweet.

Thus, results can be interpreted the following way: *Words* represent the keyword arguments according to which the top clusters produced in each hour from the first clustering step were bundled together in the second clustering step. Their assigned *weights* represent the overall importance of these specific words during the respective

time of the day. Recalling that these weights are a result of the df-idf$_t$ method, the values change with each time slot. Since I selected both the earliest and the most informative Tweet from each cluster, a maximum of two weight sets are applied for each cluster, yielding a maximum of two different weight sets for the *same* cluster. In simple terms, this means that similar talk about similar topics from different times of day can be visualized and analyzed separately.

Examples for such cluster wordcloud and Tweet combinations are shown below in figures 19, 20, 21, and 22; each together with information from news reports for the purposes of a first comparison.

Tweet text	Taksim Gezi Parkı'nda ağaçlar çatır çatır söküldü: İstanbul'da Topçu Kışlası'nın yapılması planlanan Gezi Pa... bit.ly/1417bqA
Translation	Taksim Gezi Park trees are being shredded: The construction of the Topcu Barracks at Gezi Pa… bit.ly/1417bqA
Time	2013-05-28 11:02:11
Reported event: MP Sirri Sürreya Önder joins protesters protecting the park from the bulldozers, then police evicts the protesters by force. (Reported by: NTV, Hürriyet)	

Figure 19: Combined system output and comparison to a reported event (example 1)

Tweet text	Gezi Parkı'nı 5 dakika içinde kaybedeceğiz ne yazık ki!
Translation	We will lose the Gezi Park within 5 minutes how sad!
Time	2013-05-30 02:00:07
Reported event: Nightly raid on Gezi Park; heavily armed riot police attempt to drive out the protesters with water cannons, tear gas and batons. (Report by: Hürriyet)	

Figure 20: Combined system output and comparison to a reported event (example 2)

Tweet text	Ultraslan : Parka polis sokmuyor Gfb : gezi parkini koruyor CARŞI : panzerle polis kovaliyor :)))
Translation	Galatasaray: No police gets into the park Fenerbahce: Protecting the Gezi Park BEŞIKTAŞ: Attacking the police with a tank
Time	2013-06-01 17:03:15
Reported event: Intense clashes between protesters and riot police leave many wounded in the Besiktas district of Istanbul. (Reported by: Reuters)	

Figure 21: Combined system output and comparison to a reported event (example 3)

Tweet text	3 5 çapulcu olarak Taksim Gezi Parkında eyleme devam ederken, 230'a yakın Harvard mezunu profesör havaalanında başbakanı karşılıyor.
Translation	While 3 to 5 are continuing to showcase their support in Gezi Park as looters, 230 almost-Harvard-graduate-professors are greeting the Prime Minister at the airport.
Time	2013-06-06 22:00:04
Reported event: Thousands of Erdogan's supporters greeting the returning Prime Minister at the airport with shouts "we will die for you" and "let's go crush Taksim". (Reported by: Al Jazeera)	

Figure 22: Combined system output and comparison to a reported event (example 4)

A more detailed look at the keywords in the wordcloud and the information in the Tweet in example one (figure 19) shows that the combined review of the two items already allows linkage of the information:

- Keywords: *Gezi*, *Park*, *Taksim*, *police*, *guarding*, *gas*, *struggle*, *bulldozer*, *tree*, *pull*, *cut*, *Sirri*, *Süreyya*, *Önder, nature, wipe out, Besiktas*.
- Tweet: Describes the shredding of trees in Gezi Park, and includes a link to a Hürriyet news source which had live coverage of the event.

Analogically, for example three in figure 21:

- Keywords: *Besiktas*, *gas*, *situation*, *bad*, *support*, *hospital*, *doctor*, *wait*, *wounded*, *searching*, *lack*
- Tweet: Implies heavy violence between Police and protesters in the Besiktas district.

The Reuters report included the information that people who were in urgent need of medical assistance were treated in makeshift clinics in Besiktas due to the lack of ambulances and the inavailability of doctors; such conclusions can easily be drawn from the keywords and the Tweet.

In these specific cases shown above, I concluded that the combined information from such a Tweet and wordcloud pair was sufficient to gain an overview of what was happening at the time.

In other cases, the wordcloud contained relevant keywords but the corresponding headline Tweet was insufficiently clear to provide the necessary context to allow sense-making by linkage and interpretation of the terms. Therefore, overall results remain unclear and further evaluation is required.

3.5 Evaluation

This chapter is to provide a detailed evaluation of the results returned by the applied method. Recalling the goal to determine *to which degree* events can be detected with computational methods, an evaluation approach consisting of three parts was devised in "Business Understanding":

1. Comparison of the system output to the reported events which are recorded in the tables of *major* and *non-major* events.
2. Analysis of similarities and differences between samples which were detected and samples which were not detected.
3. Analysis of the rate of informative system outputs regarding Gezi Park Uprising events.

First, I went through the Tweet-wordcloud combinations outputted by the system on a day-by-day basis and looked for a matching report in the tables of *major events* and *non-major events*. The detailed results from this analysis are recorded in the same tables in the appendix, where I added a column titled *evaluation*: Here, the events were labeled as "yes" if there was a clear match, "no" if there was none, and as "ambiguous" if there was a match with insufficient or inconsistent information. The values recorded in the

time column represent the exact time of the Tweet from the sample combination. A summary of results is found in table 15 and table 16 for the respective sets of event reports.

Label	Count	Rate
Detected	17	65 %
Not detected	3	12 %
Ambiguous	6	23 %

Table 15: Results for major events

Label	Count	Rate
Detected	5	13 %
Not detected	27	71 %
Ambiguous	6	16 %

Table 16: Results for non-major events

It is evident that the method performed better in detecting events which were previously judged to be major events, than for non-major events. If my judgement of the significance of the events to the uprising is correct, the results indicate that 65 % of major events were detected with a clear match, and 23 % were detected with a weak match, which is a total of 88%. In the other case, these rates would be a mere coincidence – but the quality of my judgement is supported by the low detection rate of non-major events.

However, weak matches would probably present insufficient information to an uninformed viewer. For a viewer with good knowledge of Turkish politics, geography, and culture, the weak matches contained enough information to gain an overview of what was happening at the time.

In order to determine the reasons for weak matches, I took a closer look at these cases. For example, the event M2 includes the information that MP Sirri Süreyya Önder joined protesters on the first day. The level of detection was marked as ambiguous because at 11 am the only indication of his involvement was the prominent appearance of each part of his name in the wordcloud. But there is no reference to the MP in the Tweet, meaning

that it is left to the interpretation of the viewer to decide if he was there or not. However, subsequent headline Tweets contained clear indications that he was physically present. So an accurate impression of the situation would have been formed over time.

M10 follows a similar pattern: The early morning march of thousands of people across Birinci Köprü (one of the bridges crossing the Bosphosus Strait) was not mentioned in any headline Tweets. However, this event seems to have been arranged via calls on social media to join the march, as the word *bridge* ("Köprü") was one of the most prominent words in the wordcloud at 1 am. Furthermore, the comparatively high density of wordclouds in the hours between 4 and 7 am indicates a greater total volume of similar Tweets than usual at around 6 am. This speaks for the performance of the method in building relevant clusters – it simply seems that the cluster size was too large to allow a reading out of the details. This could be addressed by tightening the parameter settings for the dendogram cutoff height.

A closer look at the Tweets in these clusters yielded further insights: It seems that people were not primarily tweeting about the activity of marching across the strait, but focusing instead on the sense of community and on voicing their demands. The headline Tweet at 5:01 am reads:

> *"Last night we conquered the streets, now we come back stronger and the tsunami is growing (...)!"* (reference to a wave of people)

The headline Tweet at 6:12 am reads:

> *"The crimes committed by the government so far have only been salt and pepper for the Gezi Park movement – but enough is enough!"*

The observation that people were not primarily concerned with reporting details about the situation on the ground seems to point to a conceptual problem for detecting events in a political uprising context. It could be explained by the fact that the situation during the march over the bridge was peaceful – eg. no police violence occurred at the time. Perhaps the fact that the participants of the march were surrounded by thousands of

other people with smartphones influenced their decision not to Tweet about the activity; this could have consciously or subconsciously given them the impression that the activity had already been reported by many people and thus, their Tweet would not have added value. On the other hand, it seems easier to communicate the impressive visual effect of thousands walking on a bridge ("tsunami" reference above) by sharing a picture or a video instead of text – especially while walking and trying to avoid getting trampled by a crowd. This could also be an explanation; in this case, text-based analysis would not be the right approach to detect this kind of event.

Similar reasons could also explain the *ambigious* labels in M8, M21 and M22: All of these events were comparatively calm, and they matched with dense wordclouds and Tweet headlines lacking specific information about the situation on the ground. So for the Gezi case, the conclusion can be drawn that events involving violence (or perhaps, injustice) can be detected more easily with the applied method. This could be true for social media analytics in general because it concerns the psychology of sharing and human behavior in a crowd – but that would need to be examined in another study.

M14, Erdogan calling the protesters a "handful of looters", was marked as *ambiguous* because the Tweet did not specifically mention that the Prime Minister made a statement. However, the word *çapulcu* (looter) was prominent in the vocabulary, along with the name of the Prime Minister and words translating to "saying", "dear", and "Prime Minister". So the event could have been inferred from the statement. It could also have been inferred from numerous other headline Tweets containing jokes about being a looter – for example, people started Tweeting pictures of themselves captioned with their name and the description *çapulcu*, and a video titled "Everyday I'm çapulling" emerged with a Gezi Park remake of the LMFAO song "Everyday I'm shufflin". Naturally, it was not difficult to detect viral content.

Regarding the non-major events, some of the events which were not detected (eg. S18, S19 and S27) were highly specific in their level of detail. Furthermore, they may not have been witnessed by enough people, may not have been considered exceptional or noteworthy, or the keywords describing them were not unique enough to lead to the development of a separate cluster.

In other cases issues with the reports emerged, making assessment difficult: For example, S5 describes police throwing tear gas on protesters from a helicopter in Ankara. The date of the *Reuters* report was May 31 according to the website, but I found clear indications of such an event in the headline at 12:53 on June 1, which means that either the report was falsely dated, or that the event received more attention on Twitter on June 1 than on the day it actually took place. Other conclusions are also possible – eg. it may have been a rumor.

The events S13, S23 and S24, which describe incidents like the manhandling of reporters in an office building, and the usage of a university canteen as a makeshift clinic were reported by the organizations *Amnesty International* and *Reporters without Borders*. In my judgement, the information contained in these reports was probably gathered from interviews conducted weeks after the incident – which does not mean the events did not happen, but it could be that I was looking for their traces in the wrong time slots. In other cases like S22, where the report was produced by *Marxist*, the accuracy of the reported event details should be questioned due to the bias of the media outlet – in this case, the usage of expressions like "the heroic movement of the masses of Turkey continues" indicate lack of objectivity in their report. The reported *violent clashes in Besiktas* aspect was detected; concurring with M18.

Furthermore, S12 was detected: reports of armed government supporters joining riot police to crack down on protesters in Izmir. This speaks for the usability of the system as a situational awareness tool, as this event would have constituted relevant information for personal risk assessment. Similar information was also found in headline Tweets on other protest days – indicating that warnings regarding potential security threats were shared and that the method successfully grouped them.

The list could have been explored further, but the examples and numbers provided above are sufficient to conclude that the method performed better in isolating and retrieving information about *major* events that happened during the Gezi Park Uprising, than for other kinds of events. Regarding the events in the category judged as *non-major* for the overall course of the uprising, results are not clear enough to draw any conclusions. This is due to the lack of an objective evaluation method, which would

require a reliable source of truth and a large enough test sample containing events of distinguishable categories.

Regarding the performance of the method in clustering Tweets according to Gezi Park Uprising events, the count of *true positives* describes how many headline Tweets included information which could be linked directly to specific events of the given day. Thus, it is not a measure of quality of the information retrieved (as the name suggests), but an indication of the degree to which final clusters were built based on event-specific sets of keywords – as opposed to mere "topics" or groups of random keywords. Such a rate could be calculated precisely by analyzing the purity of the each cluster vocabulary in regards to the event to which the headline Tweets refered, and then averaging the scores. However, calculating the purity of 185 wordclouds would have been too time consuming; therefore, only the information in the headline Tweets was used as a heuristic for each cluster – meaning that the resulting rate should be read as a *highly conservative* estimate.

The results, which are shown in table 17, indicate that the performance was best on the data from May 31, June 1, and June 2 respectively, reaching up to 88 % retrieval of informative clusters. On nine of thirteen days the rate of relevant headlines was above 50 %.

Day	**True positives**	**Rate**
May 28	11	29 %
May 29	5	63 %
May 30	8	50 %
May 31	31	67 %
June 1	46	88 %
June 2	21	88 %
June 3	15	44 %
June 4	8	50 %
June 5	24	52 %
June 6	20	51 %
June 7	12	40 %
June 8	0	0 %
June 9	10	50 %
June 10	0	0%

Table 17: System output scores

3.6 Review

The application of the method to the Gezi Park Uprising case was highly successful in detecting major events that occurred during the first thirteen days of the protests, showing that the Ifrim et al. method could have been used as a situational awareness tool during this uprising, and that the hierarchical clustering method also can be applied to a dataset of Tweets written in another language – in this case, requiring only minor adaptation in the preprocessing phase.

Overall, the results were satisfactory: The system slightly outperformed the *true positives* rate reported by Ifrim et al. on June 1 and 2, but it underperformed by comparison on the other days.

In several points, the system could be improved further to yield better results. This is mainly related to reduction of noise in the data set: The weighting scheme proposed by Ifrim et al. relies heavily on named entities, which in my case also included the location mentions. In total, the list of names entities produced from the original data from the preprocessing phase had 15865 unique entries, including a significant share of words falsely classified as entities due to the low accuracy of the NLP Cube module on a text body with spelling errors. Based on review of samples, I estimated that more than 30 % of the words in this list could be considered noise. This considerably impacted the weighting of terms, and subsequently, the ranking of the clusters – meaning that non-event related clusters were ranked higher than they should have been.

The lemmatization step in the preprocessing phase also produced noise: Considering that the NLP Cube had an error rate of more than 15 %, the performance of the method could most likely be enhanced further by improving the lemma module or by adding a second noise filter (eg. stop word removal). This would improve the purity of the clusters regarding specific event keywords.

On the methodological side, recurring place names were weighted very highly by the Ifrim et al. method due to the repetitive nature of the events. This was only partially constrained by the penalization scheme: Words such as “gezi”, “park” and “taksim” steadily increased in weight, sometimes gaining values in excess of 1500, amplified

additionally by the entity boost factor. At the same time, more informative words with details about what happened in these places at different times of the day were weighted at levels closer to the median (I estimate it to be in the vicinity of 15, based on samples surveyed). As a direct result, clusters which did not have these words in them were ranked lower and therefore had a lower probability of being included in the top twenty selection. Perhaps this made it easier for the method to detect *major* events which occurred in the vicinity of Taksim Square's Gezi Park.

Due to the large variance in Tweet volume across time slots it was not possible to conduct a detailed parameter analysis with distinguished settings, as both the low number of Tweets in some time slots and the high number of Tweets in others had a limiting effect on the applicability of the method. Based on experiments with different cutoff heights in sample time slots it was clear that the granularity of topic clustering within a time slot can be engineered precisely; however, only a medium setting was examined due to the use of the dividend of 3.5 to achieve similar cutoff results across all slots. This method yielded clusters of a size ranging between 10 Tweets (the defined minimum) and about 200. Raising the cutoff height would have produced fewer clusters of sizes ranging between 10 and the total number of Tweets in the time slot – which typically meant one unproportionately large cluster and a small number of little clusters. It is unclear whether such a distribution would yield good results for event detection, but my impression from test runs was that this is not the case due to greater fragmentation of topics within the large cluster. Lowering the cutoff height improved the event purity in some cases, but it did not lead to a better rate of informative content in the slots.

The novel presentation method of visualizing the cluster vocabularies using the term weights from the slots made review easier, as the most important keywords describing the respective event clusters consistently appeared in bold letters in the wordclouds.

4 Conclusion

The comparison of fifteen distinguished event-detection methods in chapter two revealed that on one hand, research in this field is already quite mature, and on the other hand, solutions to the same problem can be addressed by applying different detection approaches of varying levels of complexity. Clustering approaches have the advantage that they do not necessarily require labeled datasets for training, making them highly suitable for the explorative data mining approach.

Based on the findings of the case study, hierarchical clustering is a suitable approach for the detection of *major events* in the context of a crisis.

Combinations of clustering and classification were found to be more suitable for practical applications of event detection in the crisis context, such as early warning for emergency services. In this case, the availability of training datasets with Tweets labeled according to their reference to specific, known categories of events was found to be the main enabler of detection precision. Notably, systems have also been build which discover the categories autonomously, allowing automatic labeling of Tweets for the purpose of training a classification module which then determines if an outputted cluster presents an event or not (compare Hua et al.). Graph-partitioning methods seem to be the highly versatile in their applicability to the event detection problem, as the data can be modeled in a flexible, customized schema according to the data mining goals pursued – thus allowing the coverage of more specific use cases, such as earthquake detection.

Considering the insights from the case study, there are indications that for some types of events the text-based clustering method performed poorly due to the lack of citizen reporting about the action taking place in their immediate surroundings. It seems likely that for such cases a computer vision method would be more suitable. Therefore, methods like the deep learning approach of Won et al. for image-based event detection in social media present an interesting alternative.

4.1 Limitations

The findings in this thesis represent a major step forward for event detection in Turkish social media posts, as this was shown to be feasible for the first time (to my knowledge).

However, the case study was limited by multiple factors: For example, it was unclear to which degree the acquired data set represents the complete amount of Tweets published in the 13-day time window – both in regards to whether the Twitter advanced search page returned only a selection, and whether my own choice of hashtags to search for covered a statistically representative set of Tweets from the uprising.

The evaluation method was strongly influenced by my own judgement, as I assessed event significance of the preselected events based on my expert knowledge of Turkish politics and culture. Although this stems from having observed the political situation closely while I lived in the country for more than ten years (including the year 2013), this presents a limiting factor for the interpretability of results. For example, considering that intuition played a role in the interpretation of the system output – the wordclouds in the context of their headline Tweet – the aptitude of the system as a situational awareness tool for *uninformed* users has yet to be determined. Evaluation according to an objective, research-based set of events with clearly distinguished event categories, and using both a number of informed and uninformed test subjects for the review of the system output would have improved the research quality.

4.2 Implications for research and practice

Detecting events from social media data in a crisis context is feasible and different systems have been built to this purpose, enabling applications in humanitarian and intelligence gathering contexts. Furthermore, the election context presents a recurring (non-crisis related) application found in the research papers I surveyed. Considering that these applications all represent use cases where decisions made by actors could have negative consequences for individuals or even on a societal level, ethical questions regarding *how* such systems are actually used present an interesting object for further study of event detection. For example, the links to decision theory, policy, and

information systems could be explored in depth, as well as the criteria according to which such decision aids in crisis contexts are selected for use in a live environment.

Regarding the aspect of natural language processing in the Turkish language, part-of-speech tagging and lemmatization tools were identified as research niches with insufficient coverage to date. Regarding the case itself, it would be interesting to apply another method to the Gezi Park Uprising data for a comparison of results.

4.3 Outlook

With the basic functionality of an event detection system already in place, I envision to explore possibilities in further developing the system into a situational awareness app.

References

Acemoglu, D., & Restrepo, P. (2018). *Artificial Intelligence, Automation and Work* (Working Paper No. 24196). National Bureau of Enconomic Research.

Agarwal, N., & Liu, H. (2009). *Modeling and data mining in the blogosphere*. San Rafael, Calif.: Morgan & Claypool.

Aggarwal, C. C. (2011). An Introduction to Social Network Data Analytics. In C. C. Aggarwal (Ed.), *Social Network Data Analytics* (pp. 1–15). Boston, MA: Springer US.

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., … Jaimes, A. (2013). Sensing Trending Topics in Twitter. *IEEE Transactions on Multimedia*, *15*(6), 1268–1282.

Alitor, D., & Salomons, A. (2018). Is Automation Labor Share-Displacing? Productivity Growth, Employment, and the Labor Share. *Brookings Papers on Economic Activity*, 1–64.

Allan, J. (2002). Introduction to Topic Detection and Tracking. In J. Allan (Ed.), *Topic Detection and Tracking* (Vol. 12, pp. 1–16). Boston, MA: Springer US.

Alsaedi, N., Burnap, P., & Rana, O. (2017). Can We Predict a Riot? Disruptive Event Detection Using Twitter. *ACM Transactions on Internet Technology*, *17*(2), 1–26.

Bai, R., Wang, X., & Liao, J. (2009). Folksonomy for the Blogosphere: Blog Identification and Classification. *2009 WRI World Congress on Computer Science and Information Engineering*, 631–635. Los Angeles, California USA: IEEE.

Barbier, G., & Liu, H. (2011). Data Mining in Social Media. In C. C. Aggarwal (Ed.), *Social Network Data Analytics* (pp. 327–352). Boston, MA: Springer US.

Becker, H. (2011). *Identification and Characterization of Events in Social Media* (Doctoral Dissertation). Columbia University.

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25–71). Springer.

Bilgin, O. (2016). *Frequency effects in the processing of morphologically complex Turkish words* (Doctoral Dissertation). Boğaziçi University.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Boroș, T., Dumitrescu, S. D., & Burtica, R. (2018). NLP-Cube: End-to-end raw text processing with neural networks. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 171–179.

Cameron, M. A., Power, R., Robinson, B., & Yin, J. (2012). Emergency situation awareness from twitter for crisis management. *Proceedings of the 21st International Conference Companion on World Wide Web - WWW '12 Companion*, 695. Lyon, France: ACM Press.

D'hondt, J., Vertommen, J., Verhaegen, P.-A., Cattrysse, D., & Duflou, J. R. (2010). Pairwise-adaptive dissimilarity measure for document clustering. *Information Sciences*, *180*(12), 2341–2358.

Domeniconi, C., & Al-Razgan, M. (2009). Weighted cluster ensembles: Methods and analysis. *ACM Transactions on Knowledge Discovery from Data*, *2*(4), 1–40.

Ester, M., Kriegel, H.-P., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Kdd*, *96*(34), 226–231.

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM*, *57*(6), 74–81.

Feng, W., Zhang, C., Zhang, W., Han, J., Wang, J., Aggarwal, C., & Huang, J. (2015). STREAMCUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream. *2015 IEEE 31st International Conference on Data Engineering*, 1561–1572. IEEE.

Franch, F. (2013). (Wisdom of the Crowds) 2: 2010 UK election prediction with social media. *Journal of Information Technology & Politics*, *10*(1), 57–71.

Frangonikolopoulos, C. A., & Chapsos, I. (2012). Explaining the Role and the Impact of the Social Media in the Arab Spring. *Global Media Journal: Mediterranean Edition*, *7*(2).

Gaspar, R., Pedro, C., Panagiotopoulos, P., & Seibt, B. (2016). Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events. *Computers in Human Behavior*, *56*, 179–191.

Gayo-Avello, D., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., Schoen, H., Gloor, P., … Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research: Electronic Networking Applications and Policy*, *23*(5), 544–559.

Gruhl, D., & Guha, R. (2004). Information Diffusion Through Blogspace. *Proceedings of the 13th Conference on World Wide Web - WWW '04*, 491–501. ACM.

Hammouda, K. M., & Kamel, M. S. (2004). Efficient phrase-based document indexing for Web document clustering. *IEEE Transactions on Knowledge and Data Engineering*, *16*(10), 1279–1296.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.

Heidarian, A., & Dinneen, M. J. (2016). A Hybrid Geometric Approach for Measuring Similarity Level Among Documents and Document Clustering. *2016 IEEE*

Second International Conference on Big Data Computing Service and Applications (BigDataService), 142–151. IEEE.

Hua, T., Chen, F., Zhao, L., Lu, C.-T., & Ramakrishnan, N. (2013). STED: Semi-supervised targeted-interest event detectionin in twitter. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13*, 1466. Chicago, Illinois, USA: ACM Press.

Huang, Y., Li, Y., & Shan, J. (2018). Spatial-Temporal Event Detection from Geo-Tagged Tweets. *ISPRS International Journal of Geo-Information*, *7*(4), 150.

Ifrim, G., Shi, B., & Brigadir, I. (2014). Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering. *CEUR Workshop Proceedings*, *1150*, 33–40.

Johansson, F., Brynielsson, J., Horling, P., Malm, M., Martenson, C., Truve, S., & Rosell, M. (2011). Detecting Emergent Conflicts through Web Mining and Visualization. *2011 European Intelligence and Security Informatics Conference*, 346–353. Athens, Greece: IEEE.

Jongman, B., Wagemaker, J., Romero, B., & de Perez, E. (2015). Early flood detection for rapid humanitarian response: Harnessing near real-time satellite and Twitter signals. *ISPRS International Journal of Geo-Information*, *4*(4), 2246–2266.

Khondker, H. H. (2011). Role of the new media in the Arab Spring. *Globalizations*, *8*(5), 675–679.

Korkmaz, G., Cadena, J., Kuhlman, C. J., Marathe, A., Vullikanti, A., & Ramakrishnan, N. (2015). Combining Heterogeneous Data Sources for Civil Unrest Forecasting. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*, 258–265. Paris, France: ACM Press.

Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C.-C. (2012). Tedas: A twitter-based event detection and analysis system. *2012 IEEE 28th International Conference on Data Engineering*, 1273–1276. IEEE.

Liu, B., & Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 415–463). Boston, MA: Springer US.

Lopez, A., Coughlan de Perez, E., Bazo, J., Suarez, P., van den Hurk, B., & van Aalst, M. (2018). Bridging forecast verification and humanitarian decisions: A valuation approach for setting up action-oriented early warnings. *Weather and Climate Extremes*.

Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications*, *38*(10), 12708–12716.

Manning, C., Raghavan, P., & Schuetze, H. (2009). Introduction to Information Retrieval. *Natural Language Engineering*, *16*(1), 100–103.

Mathioudakis, M., & Koudas, N. (2010). TwitterMonitor: Trend detection over the twitter stream. *Proceedings of the 2010 International Conference on Management of Data - SIGMOD '10*, 1155. Indianapolis, Indiana, USA: ACM Press.

Maynard, D., & Funk, A. (2012). Automatic Detection of Political Opinions in Tweets. In R. García-Castro, D. Fensel, & G. Antoniou (Eds.), *The Semantic Web: ESWC 2011 Workshops* (Vol. 7117, pp. 88–99). Berlin, Heidelberg: Springer Berlin Heidelberg.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113.

Melville, P., Sindhwani, V., & Lawrence, R. (2009). Social Media Analytics: Channeling the Power of the Blogosphere for Marketing Insight. *Proceedings of the WIN*, *1*, 1–5.

Mishler, A., Wonus, K., Chambers, W., & Bloodgood, M. (2017). Filtering tweets for social unrest. *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, 17–23. IEEE.

Moon, I.-C., Kim, Y.-M., Lee, H.-J., & Oh, A. H. (2009). Temporal Issue Trend Identifications in Blogs. *2009 International Conference on Computational Science and Engineering*, 619–626. Vancouver, BC, Canada: IEEE.

Moore, G. E. (1965). Cramming More Components Onto Integrated Circuits. *Electronics*, *38*(8), 114–117.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *ArXiv Preprint ArXiv:1103.2903*. Retrieved from http://arxiv.org/abs/1103.2903

Öztürk, N., & Ayvaz, S. (2018). Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics*, *35*(1), 136–147.

Ozturkcan, S., Kasap, N., Cevik, M., & Zaman, T. (2017). An analysis of the Gezi Park social movement tweets. *Aslib Journal of Information Management*, *69*(4), 426–440.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web.* Stanford InfoLab.

Paice, C. D. (1994). An Evaluation Method for Stemming Algorithms. In B. W. Croft & C. J. van Rijsbergen (Eds.), *SIGIR '94* (pp. 42–50). London: Springer London.

Palen, L., Vieweg, S., Liu, S. B., & Hughes, A. L. (2009). Crisis in a Networked World: Features of Computer-Mediated Communication in the April 16, 2007, Virginia Tech Event. *Social Science Computer Review*, *27*(4), 467–480.

Panagiotou, N., Katakis, I., & Gunopulos, D. (2016). Detecting Events in Online Social Networks: Definitions, Trends and Challenges. In S. Michaelis, N. Piatkowski, & M. Stolpe (Eds.), *Solving Large Scale Learning Tasks. Challenges and Algorithms* (Vol. 9580, pp. 42–84). Cham: Springer International Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(October), 2825–2830.

Que, X., Checconi, F., Petrini, F., & Gunnels, J. A. (2015). Scalable community detection with the louvain algorithm. *2015 IEEE International Parallel and Distributed Processing Symposium*, 28–37. IEEE.

Ranneries, S. B., Kalør, M. E., Nielsen, S. Aa., Dalgaard, L. N., Christensen, L. D., & Kanhabua, N. (2016). Wisdom of the local crowd: Detecting local events using social media data. *Proceedings of the 8th ACM Conference on Web Science - WebSci '16*, 352–354. Hannover, Germany: ACM Press.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14–46.

Reed, J., Jiao, Y., Potok, T., Klump, B., Elmore, M., & Hurson, A. (2006). TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams. *2006 5th International Conference on Machine Learning and Applications (ICMLA'06)*, 258–263. Orlando, FL, USA: IEEE.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520.

Sadilek, A., Kautz, H., & Silenzio, V. (2012). Predicting disease transmission from geo-tagged micro-blog data. *Preceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). SentiCircles for Contextual and Conceptual Semantic Sentiment Analysis of Twitter. In V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, & A. Tordai (Eds.), *The Semantic Web: Trends and Challenges* (Vol. 8465, pp. 83–98). Cham: Springer International Publishing.

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management*, *52*(1), 5–19.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *Proceedings of the 19th International Conference on World Wide Web*, 851–860.

Scherer, M. U. (2016). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology*, *29*(2), 353.

Schulz, A., Schmidt, B., & Strufe, T. (2015). Small-Scale Incident Detection based on Microposts. *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, 3–12. Guzelyurt, Northern Cyprus: ACM Press.

Sinnott, R. W. (1984). Virtues of the Haversine. *Sky Telescope*, *68*, 159.

Social Media and Political Participation Lab. (2013). *SMaPP Data Report: A Breakout Role for Twitter? The Role of Social Media in the Turkish Protests*. Retrieved

from New York University website: https://smappnyu.org/wp-content/uploads/2018/11/turkey_data_report.pdf

Starbird, K., & Palen, L. (2012). (How) Will the Revolution be Retweeted? Information Diffusion and the 2011 Egyptian Uprising. *Proceedings of the Acm 2012 Conference on Computer Supported Cooperative Work*, 7–16.

STATISTA. (2018). *Number of social media users worldwide from 2010 to 2021 (in billions)*. Retrieved from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Steinhaus, H. (1956). Sur la division des corps materiels en parties. *Bulletin L'Académie Polonaise Des Sciences*, *4*(3), 801–804.

Subrahmanian, V. S. (Ed.). (2013). *Handbook of Computational Approaches to Counterterrorism*. New York, NY: Springer Science & Business Media.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163–173.

Varol, O., Ferrara, E., Ogan, C. L., Menczer, F., & Flammini, A. (2014). Evolution of Online User Behavior During a Social Upheaval. *Proceedings of the 2014 ACM Conference on Web Science - WebSci '14*, 81–90.

Walther, M., & Kaisser, M. (2013). Geo-spatial Event Detection in the Twitter Stream. In P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, … E. Yilmaz (Eds.), *Advances in Information Retrieval* (Vol. 7814, pp. 356–367). Berlin, Heidelberg: Springer Berlin Heidelberg.

Wang, F., & Landau, D. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letters*, *86*(10), 2050.

Weng, J., Yao, Y., Leonardi, E., & Lee, F. (2011). Event Detection in Twitter. *Proceedings of the Fifth International Conference on Weblogs and Social Media*, 22.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29–39.

Won, D., Steinert-Threlkeld, Z. C., & Joo, J. (2017). Protest Activity Detection and Perceived Violence Estimation from Social Media Images. *Proceedings of the 25th ACM International Conference on Multimedia*, 786–794.

World Bank. (2018a). *Individuals using the Internet (% of population)*. Retrieved from https://data.worldbank.org/indicator/IT.NET.USER.ZS

World Bank. (2018b). *Population, total*. Retrieved from https://data.worldbank.org/indicator/sp.pop.totl

Xie, J., & Szymanski, B. K. (2011). Community detection using a neighborhood strength driven label propagation algorithm. *2011 IEEE Network Science Workshop*, 188–195. IEEE.

Yang, S.-F., & Rayz, J. T. (2018). An Event Detection Approach Based On Twitter Hashtags. *ArXiv Preprint ArXiv:1804.11243*, 22.

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, *58*(4), 479–493.

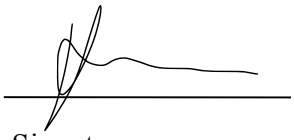
Affidavit

I hereby affirm that this Master's Thesis represents my own written work and that I have used no sources and aids other than those indicated. All passages quoted from publications or paraphrased from these sources are properly cited and attributed.

This thesis was not submitted in the same or in a substantially similar version, not even partially, to another examination board and was not published elsewhere.

August 8, 2019

Date

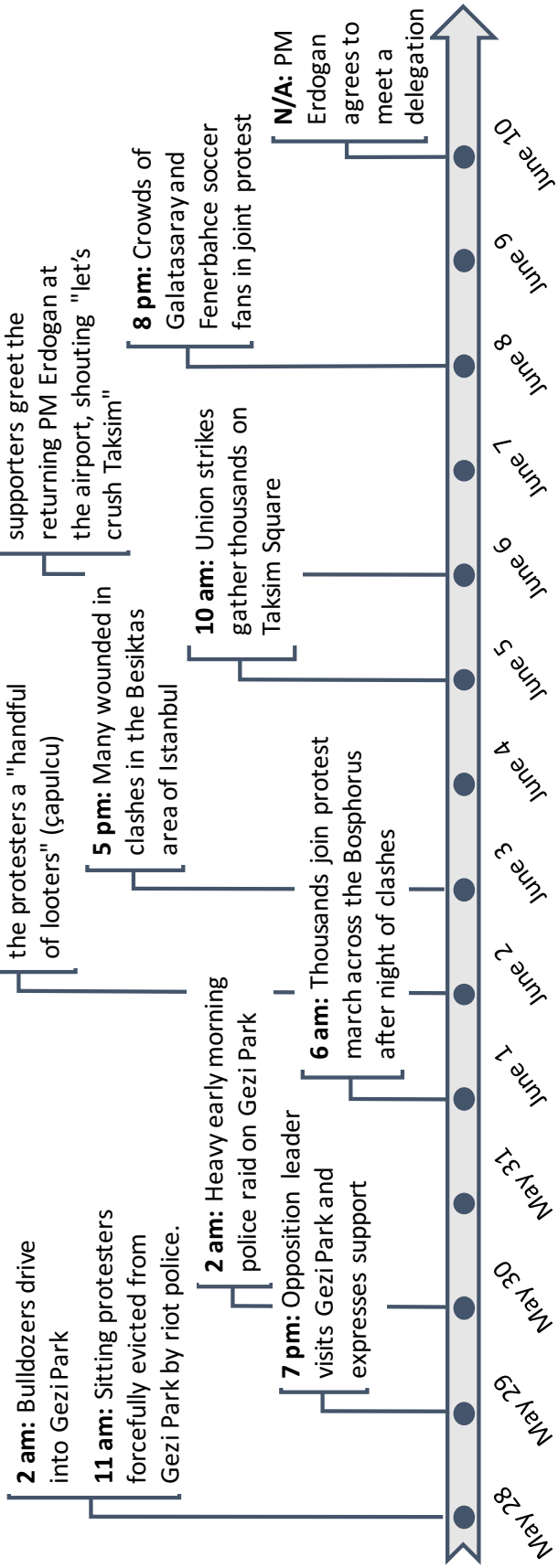


Signature

Appendix

Table of Contents

A1: Timeline of events



A2: Table of major events

Major Events					Evaluation		Link to news source
ID	**Date**	**Description**	**Source**	**Time according to report**	**Detected**	**Time**	
M1	5/28/13	About 50 people join sitting protest against the cutting of trees with a bulldozer at Gezi Park, Taksim Square – which began at 2 am	NTV	Early morning	yes	02:07	http://www.ntvmsnbc.com/id/25445552
M2	5/28/13	MP Sirri Sürreya Önder joins protesters protecting the park from the bulldozers.	Hürriyet Daily News	Early morning	ambiguous	11:02	http://www.hurriyetdailynews.com/main-opposition-leader-lends-full-support-to-protesters-occupying-taksim-gezi-park-47849
M3	5/28/13	Police evict the sitting protesters from the park using batons and tear gas	Hürriyet Daily News	Early morning	yes	10:04	http://www.hurriyetdailynews.com/main-opposition-leader-lends-full-support-to-protesters-occupying-taksim-gezi-park-47849
M4	5/29/13	MPs and celebrities join the protesters in Taksim Square	NTV	N/A	no	-	http://www.ntvmsnbc.com/id/25445854/
M5	5/29/13	Gezi Park camp is re-built; opposition leader Kilicdaroglu visits the park and expresses support	Hürriyet Daily News	N/A	yes	19:20	http://www.hurriyetdailynews.com/main-opposition-leader-lends-full-support-to-protesters-occupying-taksim-gezi-park-47849
M6	5/30/13	Organized police raid on re-established Gezi Park camp: Police using water cannons and tear gas,	Hürriyet	Before dawn	yes	02:00	http://hurarsiv.hurriyet.com.tr/goster/haber.aspx?id=23402207&tarih=2013-05-30

		but don't succeed in taking the park					
M7	5/31/13	Heavy early morning raid on Gezi Park, protesters cleared out completely	Turkish Weekly	Early morning	yes	02:02	http://www.turkishweekly.net/news/151041/-occupy-taksim-grows-in-spite-of-crackdown.html
M8	5/31/13	More than 10 000 protesters gather in Istiklal Avenue in front of Divan Hotel, following calls on Social Media	The Telegraph	Afternoon and evening	ambiguous	01:33 (next day)	https://www.telegraph.co.uk/news/worldnews/europe/turkey/10092490/Tear-gas-fired-at-protesters-in-Istanbul.html
M9	5/31/13	Clashes on Taksim Square go on throughout the day	France24	All day	yes (multiple)	13:39	http://www.france24.com/en/20130531-dozens-injured-istanbul-protest-turkey-police-clashes-demonstrators
M10	6/1/13	Thousands walk across the first Bosphorus bridge (Birinci Köprü) in protest	BBC	06:00	ambiguous	01:00	https://www.bbc.co.uk/news/world-europe-22739423
M11	6/1/13	Tear gas fired at protesters in Taksim Square	NTV	N/A	yes	12:00	http://www.ntvmsnbc.com/id/25446372/
M12	6/1/13	Police forces withdraw from the square early on Sunday, protesters celebrating	The Telegraph	Early morning	yes	07:14	https://www.telegraph.co.uk/news/worldnews/europe/turkey/10093974/Turkey-protesters-celebrate-after-police-leave-Istanbul-square.html
M13	6/2/13	Following nightly clashes, police withdraw from Taksim Square and	Hürriyet Daily News	Morning	yes	09:08	http://www.hurriyetdailynews.com/protesters-clean-taksim-area-after-police-withdrawal.aspx?pageID=238&nID=48029&NewsCatID=341

		protesters gather to clean up the debris					
M14	6/2/13	PM Erdoğan calls the protesters "a handful of looters" (çapulcu)	Radikal	N/A	ambiguous	13:33	http://www.radikal.com.tr/politika/basbakan_erdogan_biz_birkac_capulcunun_yaptiklarini_yapmayiz-1136875
M15	6/2/13	Heavy clashes between riot police and protesters in Izmir leave many store-fronts damaged	LiveLeak	Afternoon	yes	16:05	http://www.liveleak.com/view?i=5ba_1370204264
M16	6/2/13	President of the Turkish Bar Association publicly speaks up against the excessive police violence, tear gassing of patients and beating of doctors in makeshift infirmary	Son Dakika	22:00	no	-	http://www.sondakika.com/haber/haber-tbb-baskani-feyzioglu-mulkiyeliler-birligi-ne-4692334/
M17	6/3/13	Police violently storms barricades in the Dolmabahce area, Istanbul	LiveLeak	N/A	yes	05:00	http://www.liveleak.com/view?i=b08_1370215088
M18	6/3/13	Intense clashes between protesters and riot police leave many wounded in the Beşiktaş district of Istanbul	Daily Mail	N/A	yes	17:03	https://www.dailymail.co.uk/news/article-2334989/Turkey-protests-Twenty-year-old-protester-KILLED-Turkey-taxi-mows-demonstrators-fourth-day-violence-growing-Islamic-influence.html
M19	6/4/13	Tens of thousands gathering at Taksim Square in what is	HaberBiz	All day and night	yes (multiple)	19:00	http://www.haberbiz.com/turkiye_bugun_siyah_giyecek-guncel-haber-46899.html

		described as a festival atmosphere					
M20	6/5/13	Union strikes across the country: Taksim Square gathers the largest number of people to date. Crowds of protesters and AKP supporters seen attacking each other.	Hürriyet, BBC	Afternoon	yes	10:00	http://www.hurriyetdailynews.com/crowd-attacks-supporters-of-gezi-park-protests-in-erdogans-homeland.aspx?pageID=517&nID=48309&NewsCatID=341
M21	6/5/13	10 000 union strikers protesting in Kizilay Square, Ankara	Son Dakika	Afternoon	ambiguous	17:02	http://www.sondakika.com/haber/haber-kesk-ve-disk-in-kizilay-mitingi-suruyor-4701214/
M22	6/5/13	Protest in Ankara gathers thousands at Kizilay Square but the situation was mostly peaceful	CNN	Afternoon	ambiguous (multiple)	18:05	https://edition.cnn.com/2013/06/05/world/meast/turkey-woman-in-red/index.html
M23	6/6/13	Thousands remain in Taksim Square during the night. Because it was an Islamic holy night, there were no heavy-handed clashes with police.	Time	Night	yes	23:00	http://world.time.com/2013/06/05/live-from-occupied-gezi-park-in-istanbul-a-new-turkish-protest-movement-is-born/

M24	6/6/13	AKP party runs a SMS campaign and organizes buses to gather thousands of supporters at Ataturk Airport, where PM Erdogan is set to arrive, returning from Morocco. The crowd of around 10 000 people chants "We will die for you, Erdogan", "Let's go crush them all", "Lets go crush Taksim".	Al Jazeera	Evening	yes	21:00	http://www.aljazeera.com/news/europe/2013/06/20136705734678575.html
M25	6/8/13	Protests continue on Taksim square with the support of two large football clubs. Fenerbahce and Galatasaray supporters collaborate to place flares on the top of the Atatürk Cultural Center.	Hürriyet	Evening	yes	20:00	http://www.hurriyetdailynews.com/team-work-of-united-ultras-set-taksim-on-fire.aspx?pageID=238&nID=48464&NewsCatID=341
M26	6/10/13	PM Erdogan agrees to meet a representation of protesters	BBC	N/A	no	-	https://www.bbc.co.uk/news/world-europe-22844461

A3: Table of non-major events

Non-Major Events					Evaluation		Link to news source
ID	**Date**	**Description**	**Source**	**Time according to report**	**Detected**	**Time**	
S1	5/28/13	Member of Parliament Sirri Önder physically stands in front of moving bulldozer	NTV	Morning	ambiguous	11:02	http://www.ntvmsnbc.com /id/25445552
S2	5/28/13	Police is photographed spraying tear gas into the face of a female protester wearing a red dress (goes viral on Twitter and in the news)	Washington Post	Morning	no	-	https://www.washingtonpost.com/ blogs/worldviews/wp/2013/06/03/ the-photo-that-encapsulates-turkeys-protests-and-the-severe-police-crackdown/
S3	5/31/13	Protesters build barricades on Taksim Square	Turkish Weekly	N/A	no	-	http://www.turkishweekly.net/news/ 151041/-occupy-taksim-grows-in -spite-of-crackdown.html
S4	5/31/13	MP Sirri Önder is injured by flying gas canister	New York Times	N/A	no	-	https://www.nytimes.com/2013/06/01 / world/europe/police-attack-protesters-in -istanbuls-taksim-square.html?pagewanted=all
S5	5/31/13	Tear gas is fired at protesters from a helicopter in Ankara	Reuters	May 31	ambiguous	12:53 (June 1)	https://www.reuters.com/article/2013/ 06/01/us-turkey-protests-idUSBRE94U0J9201 30601

S6	5/31/13	Police are seen chasing demonstrators into shops with electric shock batons in Ankara	Reuters	N/A	no	-	https://www.reuters.com/article/2013/06/01/us-turkey-protests-idUSBRE94U0J920130601
S7	6/1/13	Cazerolazo: People in urban neighborhoods of Istanbul stand on their balconies, banging pots in protest	BBC	N/A	yes	18:02	https://www.bbc.co.uk/news/world-europe-22739423
S8	6/1/13	"TOMA" (armored water cannon vehicle) runs over protester in Ankara	MyNet	N/A	no	-	https://www.mynet.com/ankarada-bir-genc-tomanin-altinda-kaldi-110100699882
S9	6/1/13	First reports that Agent Orange chemical is used in water cannons in Istanbul	CNN	N/A	no	-	https://web.archive.org/web/20130602012019/http://ireport.cnn.com/docs/DOC-980610
S10	6/1/13	Armored police truck hits protester in Istanbul	Reuters	N/A	no	-	http://uk.reuters.com/article/2013/06/02/uk-turkey-protests-idUKBRE94U0JA20130602
S11	6/2/13	Birinci Köprü bridge in Istanbul is blocked by police to prevent march to Taksim	Al Jazeera	N/A	no	-	http://blogs.aljazeera.com/liveblog/topic/turkey-protests-20176
S12	6/2/13	Reports of armed AKP supporters joining riot police to crack down on protesters in Izmir	Al Jazeera	N/A	yes	23:09	http://www.aljazeera.com/news/europe/2013/06/201362234021816855.html

S13	6/2/13	At 8pm, a makeshift health clinic operating in the entrance floor of Kızılay Shopping Centre (Istanbul) is tear gassed by police during treatment of the injured	Amnesty International	20:00	no	-	https://www.amnesty.org/download/Documents/12000/eur440222013en.pdf
S14	6/2/13	Protesters in Istanbul hijack a catterpillar and try to use it against police positions. Later the bulldozer is set on fire.	Al Jazeera	At night	no	-	http://www.aljazeera.com/news/europe/2013/06/201362234021816855.html
S15	6/2/13	Reports of police firing tear gas into homes.	Hürriyet	N/A	no	-	http://www.hurriyetdailynews.com/timeline-of-gezi-park-protests--48321
S16	6/3/13	Protesters in Izmir set fire to the local AKP headquarters	Daily Mail	N/A	no	-	https://www.dailymail.co.uk/news/article-2335557/Turkey-protests-Government-says-sorry-protesters-desperate-bid-quell-days-violence-killed.html
S17	6/3/13	Riot police firing tear gas inside the Bahcesehir University campus.	LiveLeak	N/A	ambiguous	09:02	http://www.liveleak.com/view?i=c9c_1370207727
S18	6/3/13	Beyoglu businesses and Dolmabahce Mosque used as makeshift hospitals and shelters for wounded protesters	Haberler	N/A	no	-	http://en.haberler.com/beyoglu-businesses-aid-protesters-tourists-in-hour123456789-278807/

S19	6/3/13	Protesters and tourists with severe injuries are treated in a Besiktas Mosque due to the inavailability of ambulances	Radio1	N/A	no	-	https://web.archive.org/web/20160304092249/http://www.radio1.be/programmas/de-ochtend/krijgt-turkije-zijn-eigen-arabische-lente
S20	6/3/13	Shangri-La Hotel used as a makeshift hospital	BBC	All day	no	-	https://www.bbc.co.uk/news/world-europe-22754348
S21	6/3/13	AKP supporters beat up protesters in Izmir	CNN	N/A	no	-	http://ireport.cnn.com/docs/DOC-981187?ref=feeds%2Flatest
S22	6/3/13	Heavy clashes in the Besiktas district of Istanbul, with truck drivers joining to shield protesters from the water cannons	Marxist	Afternoon	ambiguous	18:00	http://www.marxist.com/brutal-suppression-leaves-3-dead-urgent-solidarity-needed.htm
S23	6/3/13	Ankara office of Sol newspaper raided by riot police, reporters manhandled and gassed inside the building	Reporters without borders	At night	no	-	http://en.rsf.org/turkey-occupy-gezi-protests-lead-to-wave-06-06-2013,44732.html
S24	6/3/13	During violent clashes, the Beşiktaş University's cafeteria was turned into a makeshift infirmary.	Amnesty International	N/A	no	-	https://www.amnesty.org/download/Documents/12000/eur440222013en.pdf

S25	6/3/13	More than 100 000 people wearing black on Monday, in response to a Facebook event called "Black Monday"	HaberBiz	N/A	no	-	http://www.haberbiz.com/turkiye_bugun_siyah_giyecek-guncel-haber-46899.html
S26	6/4/13	In Dolmabahçe, police is seen firing teargas grenades at a wounded protester lying on the ground	Al Jazeera	N/A	no	-	http://blogs.aljazeera.com/topic/turkey-protests/photos-show-riot-police-aiming-teargas-wounded-protester
S27	6/4/13	Protesters build fully operational kitchen, first aid clinic and other infrastructure in Taksim Square	The Week	All day	no	-	http://theweek.com/article/index/245072/dispatch-from-istanbul-occupy-gezi-park-digs-in
S28	6/4/13	Protesters build a makeshift library offering “forbidden” literature in Taksim Square	Hürriyet	All day	ambiguous	02:00 (different day)	http://www.hurriyetdailynews.com/publishing-houses-to-unite-in-gezi-park-to-distribute-major-resistance-material-books.aspx?pageID=238&nID=48234&NewsCatID=341
S29	6/4/13	Thousands of schoolchildren skip school in Ankara in protest	Nos	N/A	no	-	http://nos.nl/audio/514187-joris-van-de-kerkhof-demonstraties-in-ankara.html

S30	6/4/13	Reports of a young Kurdish demonstrator shot in the head and killed in the province of Tunceli	Marxist	N/A	ambiguous	21:00	http://www.marxist.com/brutal-suppression-leaves-3-dead-urgent-solidarity-needed.htm
S31	6/4/13	Violence in Antakya and Adana. In Tunceli, the military steps in.	Huffington Post	N/A	no	-	http://www.huffingtonpost.com/2013/06/09/erdogan-condemns-protests-hundful-looters_n_3411536.html
S32	6/4/13	Riot police in Adana using tear gas in a hospital - descriptions of "Gaza-like" scenes.	Amnesty International	N/A	no	-	https://www.amnesty.org/download/Documents/12000/eur440222013en.pdf
S33	6/4/13	16 people including a teenage girl arrested in Izmir for calling for further protests on social media	Radikal	N/A	no	-	http://www.radikal.com.tr/turkiye/izmirde_halki_isyana_tesvik_baskinlari_16_gozalti-1136298
S34	6/5/13	Cybercrimes police division in Izmir raids 38 locations, arresting 24 people who are accused of "using Twitter to urge people to come to the protests."	Milliyet	N/A	no	-	http://gundem.milliyet.com.tr/izmir-de-sosyal-medya-operasyonu-/gundem/detay/1718776/default.htm
S35	6/5/13	AKP supporters violently attack a group of protesters in Rize. Police steps in as a young girl is nearly trampled to death.	Hürriyet	N/A	yes	20:00	http://www.hurriyetdailynews.com/crowd-attacks-supporters-of-gezi-park-protests-in-erdogans-homeland.aspx?pageID=517&nID=48309&NewsCatID=341

S36	6/5/13	In Antakya in the evening thousands hold a candle-lit march in commemoration of Abdullah Cömert, a 22-year-old demonstrator who was killed in the clashes on June 3	Haaretz	N/A	no	-	http://www.haaretz.com/news/middle-east/istanbul-feels-like-a-carnival-but-the-protests-are-violent-in-turkey-s-provinces.premium-1.528039
S37	6/7/13	A group calling themselves "Anti-capitalist Muslims" prays in Gezi Park in honor of Abdullah Cömert, with many sit-in protesters gathered.	Haberler	At night	yes	01:00	http://en.haberler.com/sit-in-protest-at-gezi-park-continues-without123456789-279232/
S38	6/8/13	PM Erdoğan calls the protesters "a handful of looters" again (çapulcu), angering many	Radikal	N/A	yes	02:00	http://www.radikal.com.tr/politika/basbakan_erdogan_biz_birkac_capulcunun_yaptiklarini_yapmayiz-1136875